When F1 Is Not Enough: Structure-Aware Evaluation for Emotion Classification with the Plutchik Score

Anonymous ACL submission

Abstract

Standard metrics for emotion classification, such as macro-F1 and Jaccard, treat every misclassification as equally costly and saturate on established benchmarks. We propose the Plutchik score, a continuous similarity-based evaluation grounded in Plutchik's well-known wheel of emotions. It awards partial credit for predictions psychologically adjacent to the gold label and penalises predictions opposite to it. We pair the score with a Plutchik-aware training apparatus and apply both to four English emotion corpora (GoEmotions, SemEval-2018 E-c, CancerEmo, GoodNewsEveryone) using a single RoBERTa-based architecture and twelve training configurations. On GoEmotions, the macro-F1 spread across configurations is less than one point while the Plutchik spread is about 3.7 times wider, exposing differences that F1 hides. We also reproduce SpanEmo, the strongest accessible prior model on SemEval-2018: the two systems are within 0.014 Jaccard but differ by roughly 15 percentage points on the Plutchik score. We position the Plutchik score as a complement to F1 and Jaccard for multi-label emotion classification with structured label spaces, demonstrating that the choice of evaluation metric is not neutral.

1 Introduction

Emotion classification in text supports applications across social media analysis, dialogue systems, and affective computing (Setyoningrum et al., 2025). State-of-the-art systems use transformer encoders such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). They are evaluated almost exclusively with macro-F1 or Jaccard (Hinojosa Lee et al., 2024). These metrics treat all misclassifications as equally costly (Marwah et al., 2024). However, affective psychology models emotions as a structured space where some confusions are minor and others severe (Russell, 1980). For instance, because *happy* and *elated* occupy adjacent coordinates in this two-dimensional space, confusing them is a minor error, whereas mistaking *happy* for its spatial opposite, *sad*, represents a severe structural failure. These two limitations stem from the fact that on established multi-label emotion benchmarks F1 measures not only treat all errors as equal but also stagnate (with distinct training procedures yielding scores within sampling noise) (Card et al., 2020).

The systemic cost of treating all negative emotional signals as equal is concrete. In automated moderation pipelines, misclassifying complex expressions of grief as generic 'harmful' sentiment leads to the unwarranted removal of memorial posts, automating the erasure of public mourning (Hossain, 2025). In replies to news on social media, false stories have been shown to correlate with fear, disgust, and surprise, while true stories correlate with joy, trust, anticipation, and sadness (Vosoughi et al., 2018). Under that profile, mistaking fear for surprise leaves the inferred truth signal unchanged, whereas mistaking fear for joy reverses it. However, macro-F1 these errors are interchangeable. Against this background our main **research question** is: *how to improve emotion classification beyond F1 measures using results from psychology?*

To answer this question, we propose the **Plutchik score**, a continuous similarity-based score grounded in Plutchik's wheel of emotions (Plutchik, 1991). The choice of this model is rooted in the fact that Plutchik serves as basis for the state of the art in psychology regarding emotional models. Please note that the contribution of this paper is a new evaluation metric, not a new model. We position the Plutchik score as a complement to F1 and Jaccard, not a substitute: current metrics measure correctness, Plutchik measures structural plausibility, and the two surface different aspects of model behaviour. We apply it to four English emotion corpora (GoEmotions, SemEval-2018 E-c, CancerEmo, GoodNewsEveryone) across twelve training configurations of a single architecture, and repro-

duce one accessible SOTA model (SpanEmo) on SemEval. Across these experiments, the Plutchik score exposes systematic ranking disagreements with F1, including a 15-point gap on Plutchik between our RoBERTa-EarlyFusion system and SpanEmo at matched Jaccard. This is a significant result because it shows that even the state of the art models for emotion classification do not take into account the structural relations between the emotions as exhibited by psychology. This result thus paves the way for future research that should consider semantic psychologically-inspired scores beyond arithmetic measures.

The paper is structured as follows: Section 2 reviews related work on structure-aware evaluation, dimensional models of emotion, and existing emotion benchmarks. Section 3 defines the Plutchik score. Section 4 presents the training apparatus we pair with the score. Section 5 reports experimental results. Section 6 discusses what the score reveals. Section 7 concludes. Section 8 lists limitations.

2 Related Work

2.1 Structure-Aware Evaluation

Our motivation for a structure-aware evaluation is not unique to emotion classification. Other fields have moved away from purely point-wise correctness when the label space carries structure. In computer vision, segmentation evaluation increasingly complements pixel-level IoU with boundary-aware variants such as Boundary IoU (Cheng et al., 2021), which weights errors by their distance from the true object boundary. In graph neural networks, explainability work values models that surface interpretable rules even at modest cost in raw accuracy (Morris et al., 2024). In cybersecurity, cost-sensitive alternatives to F1 such as Cscore (Marwah et al., 2024) incorporate domain-specific error costs that the harmonic mean of precision and recall cannot capture. The unifying principle is that when the label space carries structure or when error costs are asymmetric, evaluation should reflect this. Our work extends this principle to emotion classification: the Plutchik wheel provides the structural prior, and the score we propose in Section 3 operationalises it.

2.2 Benchmark Saturation in NLP Evaluation

Static benchmarks struggle to distinguish genuine generalization from sampling noise, and over-reliance on a single isolated metric hides robustness, out-of-distribution behavior, and the structure of model errors. Card et al. (2020) characterize this problem statistically: on several tasks in the popular GLUE benchmark (notably MRPC and SST-2), test sets are now too small for most attempted SOTA comparisons to be adequately powered, meaning typical reported improvements fall below the minimum detectable effect size of the benchmark itself. Bowman and Dahl (2021) go further: they argue that NLU evaluation is broken and propose four criteria that benchmarks must satisfy, observing that current benchmarks fail these criteria and that biased systems can score highly without genuinely improving on the underlying task. Raji et al. (2021) add a construct-validity critique: a small set of dominant benchmarks is treated as a stand-in for general progress in AI, and state-of-the-art performance on these benchmarks is read as indicative of capabilities they were never designed or validated to measure. We position the Plutchik score as one response to this problem. Rather than proposing a harder benchmark, we propose a complementary axis of measurement on existing benchmarks: a score that captures the structure of errors rather than only their count. Where F1 or similar metric has saturated, Plutchik can still discriminate between training configurations, as Section 5.1 shows on GoEmotions.

2.3 Evaluation in Emotion Classification

We survey the evaluation metrics used by recent emotion classification models, drawing on leading NLP venues (ACL, EMNLP, COLING, etc.) from 2017 to 2024 using the keyword *emotion*, following the methodology of Lecourt et al. (2026). After screening for papers proposing emotion classification models, we retained 24 publications (Table 1). We then catalogued the evaluation metrics they report. The landscape is narrow. The majority of the 24 papers report a variant of F1: weighted, macro, or micro. A handful use Jaccard accuracy, plain accuracy, or area-under-the-curve scores. None report a semantically grounded metric. Every paper treats a confusion between adjacent emotions identically to a confusion between opposite emotions. At the risk of repeating ourselves let us highlight again that we position the Plutchik score as a complement to the current scores rather than a replacement: while existing measures check whether predictions match the gold label, Plutchik measures how far the wrong predictions are from the gold label on a psychologically grounded geometry.

Model	Year	Corpus type	Evaluation Metrics
Xu et al. (2024)	2024	Dialogue	Weighted F1, Micro F1
Ruan and Li (2023)	2023	Snippets	Micro F1, Macro F1, Jaccard accuracy
Gao et al. (2024)	2024	Dialogue	Weighted F1, Micro F1
Xiong et al. (2023)	2023	Dialogue	Weighted F1
Hu et al. (2021)	2021	Dialogue	Accuracy, Weighted F1, Macro F1
Shen et al. (2021)	2021	Dialogue	Weighted F1, Micro F1
Ju et al. (2021)	2021	Snippets	Micro F1, Macro F1, Micro AUC, Macro AUC
Zhang et al. (2023a)	2023	Dialogue	Weighted F1, Micro F1
Xu et al. (2023)	2023	Dialogue	Weighted F1, Micro F1
Sosea and Caragea (2021)	2021	Snippets	Average F1
Abdul-Mageed and Ungar (2017)	2017	Snippets	F-score
Wang et al. (2024)	2024	Dialogue	Weighted F1
Hou et al. (2023)	2023	Dialogue	Weighted F1, Micro F1
Mao et al. (2023)	2023	Dialogue	Micro F1
Wu et al. (2024)	2024	Dialogue	ACC, UAR, ECE, MCE, AUROC, AUPRC, NLL
Lim and Cheong (2024)	2024	Snippets	Macro F1, Micro F1
Upadhyaya et al. (2023)	2023	Snippets	Macro F1, Precision, Recall
Chen et al. (2023)	2023	Snippets	Weighted F1
Zhang et al. (2023b)	2023	Dialogue	Weighted F1, Micro F1
Ong et al. (2024)	2024	Dialogue	Macro F1
Zhang et al. (2019)	2019	Dialogue	Weighted F1
Zhao et al. (2022)	2022	Dialogue	Weighted F1
Plaza-del Arco et al. (2022)	2022	Snippets	Macro F1
Zhu et al. (2021)	2021	Dialogue	Micro F1, Macro F1, Weighted F1

Table 1: The 24 emotion classification models retained from our survey of ACL, EMNLP, IJCAI and COLING (2017–2024). None report a semantically grounded metric.

3 The Plutchik Score

We define the score in two steps. First, a similarity function between any two emotion labels. Second, a scoring rule that aggregates this similarity over a dataset.

3.1 Discrete and Dimensional Emotion Models

Affective psychology distinguishes two families of emotion models. Categorical models posit a small set of discrete primaries with no internal structure. Ekman and Keltner (1970) is the canonical example, with six basic emotions. Dimensional models instead place emotions in a continuous space. Russell's circumplex (Russell, 1980) uses valence and arousal. The PAD model (Mehrabian and Russell, 1974) adds dominance.

Plutchik's framework combines elements of both (Plutchik, 1982). It begins with eight discrete primaries: fear, anger, joy, sadness, trust, disgust, anticipation, and surprise. Each is linked to a biological adaptive function. Unlike the labels in purely categorical models, these primaries are not free-floating. In *The Emotions* (Plutchik, 1991), they are arranged on a wheel whose geometry encodes continuous structure. Position around the circle encodes similarity, while position across the circle encodes opposition, and the distance from the centre encodes intensity.

This geometry is what our score exploits. The Plutchik score does not depend on the eight primaries being the right labels. It depends on the wheel's geometry being approximately correct: similarity, opposition, and intensity. Other dimensional frameworks (Russell, PAD) could in principle yield their metric, but the idea of exploiting the geometry of the relations between emotions stays the same. We use Plutchik in this work for three reasons. First, among dimensional models, it is the most frequently referenced in NLP emotion work and it is the conceptual basis for emotional models used throughout the literature. Second, three of our four corpora (SemEval-2018, CancerEmo, GoodNewsEveryone) use label sets derived directly from Plutchik's primaries, so the geometry is not imposed externally on those benchmarks. Third, Lecourt et al. (2025) establish empirically that Plutchik's primaries align with the labels of widely used English emotion benchmarks, including all four corpora we evaluate on.

3.2 Similarity Function

We index the eight Plutchik primaries on the wheel introduced above as positions 0 to 7. *Joy* is at 0, *trust* at 1, *fear* at 2, and so on around the wheel. Multi-label benchmarks like SemEval-2018 add three dyads (*love*, *optimism*, *pessimism*). We place each dyad at the midpoint between its constituent primaries. For instance, *love* is between *joy* (position 0) and *trust* (position 1), so we put it at 0.5.

For any two emotions a and b with positions π_a and π_b, we compute the circular distance:

$$d(a,b) \;=\; \min\big(\,|\pi_a - \pi_b|,\;\; 8 - |\pi_a - \pi_b|\,\big).$$

The minimum here is what makes the wheel circular: the distance between *joy* (position 0) and *anticipation* (position 7) is 1, not 7, because we go the short way around. The similarity is then a linear function of the distance:

$$\mathrm{sim}(a,b) \;=\; 1 - 0.5 \cdot d(a,b).$$

This gives $\mathrm{sim} = 1$ for identical emotions, 0.5 for adjacent emotions on the wheel, 0 for emotions 90 degrees apart, and -1 for diametrically opposed emotions.

Mapping datasets onto the wheel. Three of our four datasets (SemEval-2018, CancerEmo, GoodNewsEveryone) use labels that lie directly on the wheel—Plutchik primaries or dyads—and require no special treatment. GoEmotions is the exception: its 28 fine-grained labels do not sit at wheel

positions and require a compound mapping that we describe next. The mapping approach generalises to any label inventory whose entries can be decomposed into $(\textit{axis}, \textit{intensity})$ components.

Compound labels for GoEmotions. We map each GoEmotions label to one or more components $(\textit{axis}, \textit{intensity})$, where $\textit{axis} \in \{0, \ldots, 7\}$ identifies a Plutchik primary and $\textit{intensity} \in \{1, 2, 3\}$ encodes the strength of the component. For example, *admiration* maps to $\{(\text{trust}, 3)\}$, *love* maps to $\{(\text{joy}, 2), (\text{trust}, 2)\}$, and *remorse* maps to $\{(\text{sadness}, 2), (\text{disgust}, 2)\}$. Between two components, similarity reuses the wheel's circular distance with an intensity penalty when the components share an axis: $1 - 0.1 \cdot |\Delta \textit{intensity}|$ if $d = 0$, and the stepped values $\{0.5, 0, -0.5, -1\}$ for $d \in \{1, 2, 3, 4\}$. Between two compound labels a and b, the overall similarity is the average of the two directional best-match aggregates: each component of a is matched against its best counterpart in b, those scores are averaged, the same is done in the reverse direction, and the two averages are themselves averaged. This compound similarity is used for GoEmotions only.

3.3 Scoring Rule

We compute the Plutchik score in two stages: first per emotion class across instances, then a macro-average across emotion classes.

For multi-label problems, let Y_i be the set of true labels for instance i and $\hat{Y}_i$ the set of predicted labels. For each emotion class c, let $\mathcal{I}_c = \{i : c \in Y_i\}$ denote the instances for which c is a true label. The per-class Plutchik score is the average best-match similarity over those instances:

$$P_c \;=\; \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \max_{j \in \hat{Y}_i} \operatorname{sim}(c, j).$$

If an instance predicts no labels ($\hat{Y}_i = \emptyset$), its contribution to P_c is 0.

The overall Plutchik score is the macro-average over the emotion classes present in the test set:

$$P \;=\; \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} P_c, \qquad \mathcal{C} = \{c : \mathcal{I}_c \neq \emptyset\}.$$

For single-label problems, $\hat{Y}_i$ is a singleton $\{\hat{y}_i\}$ and the per-class score reduces to the mean similarity $\operatorname{sim}(c, \hat{y}_i)$ over instances for which c is the true label. The dataset score remains the macro-average across classes.

In the experiments that follow, we compare the Plutchik score against macro-F1 and sample-wise Jaccard, the two most used standard metrics used in the multi-label emotion classification literature surveyed in Section 2.

4 Methodology

We use RoBERTa-EarlyFusion (Lecourt et al., 2026) as the architecture across all experiments. The choice is not central to the paper's contribution: the architecture serves as a fixed testbed in which the loss function is the only variable across our twelve configurations. We select RoBERTa-EarlyFusion because it is among the strongest published RoBERTa-based models on GoEmotions, providing a sufficiently strong baseline at which to demonstrate F1 saturation; any comparably strong transformer baseline would in principle serve the same role.

The model is a RoBERTa-base encoder augmented with General Inquirer features at the embedding layer. The General Inquirer (GI) is a content-analysis lexicon developed by Stone et al. (1966) that maps English words to hand-curated psychological and semantic categories. Each word in the lexicon receives a sparse binary vector over these categories. RoBERTa-EarlyFusion projects this vector to the sub-word embedding dimensionality and gates it before summation, letting the encoder consult the lexicon at the input layer without disturbing pretrained weights. Details are in the cited paper. The overall architecture does not change across experiments; we vary only the loss function.

4.1 Datasets

We apply our approach to four English emotion corpora summarized in Table 2. The four corpora span a range of label-set sizes (8 to 28), annotation protocols (multi-label, per-emotion binary, single-label), and sources (Reddit, Twitter, news headlines, online cancer-survivor forums). GoEmotions, the largest corpus with the densest label space, is our main experiment: this combination of dense multi-label space and large sample size is exactly the regime where macro-F1 saturates and the Plutchik score has the most room to discriminate, as our GoEmotions results demonstrate. The other three corpora serve as secondary checks of how the score behaves across label-set sizes and annotation protocols.

Title	Year	Source	Domain	Size	# of labels
GoEmotions (Demszky et al., 2020)	2020	Reddit	Generic	58,009 comments	28
SemEval-2018 E-c (Mohammad et al., 2018)	2018	Twitter	Generic	10,983 tweets	11
CancerEmo (Sosea and Caragea, 2020)	2020	Cancer Survivors Network	Health	12,031 comments	8
GoodNewsEveryone (Bostan et al., 2019)	2019	News Headlines	Generic	5,000 headlines	16

Table 2: Datasets used in this paper.

4.2 Soft Target Construction

We construct soft targets that interpolate between the one-hot label vector and a Plutchik-weighted version of it, then apply the resulting soft target to a base loss. Let $\mathbf{y}$ be the binary target vector and $\mathbf{P}$ the $C \times C$ Plutchik similarity matrix. We consider three variants. In all three, the hyperparameter $\alpha \in [0, 1]$ controls how much soft mass is added: at $\alpha = 0$ the target is one-hot, and increasing α moves the target towards the Plutchik-weighted version.

Semantic Label Smoothing (SLS). Standard label smoothing (Szegedy et al., 2016) replaces a one-hot target with a uniform mix. We replace it with a Plutchik-weighted mix:

$$\tilde{\mathbf{y}} = (1 - \alpha)\,\mathbf{y} + \alpha\,\mathrm{clip}(\mathbf{P}\mathbf{y},\, 0,\, 1).$$

We clip negative entries to zero so the result $\tilde{\mathbf{y}}$ remains a valid soft target, to which the standard binary cross-entropy loss is then applied in place of $\mathbf{y}$.

Soft Jaccard Loss (SJL). Jaccard is not differentiable. We use the Lovász hinge surrogate (Berman et al., 2018), which provides a differentiable approximation that converges to Jaccard in expectation, applied to the same Plutchik-weighted target $\tilde{\mathbf{y}}$ as in SLS. The result is a single loss term that simultaneously rewards exact overlap and partial credit for psychologically adjacent predictions.

SJL with Negative Targets (SJL-neg) A natural extension of SJL is to keep the negative entries in $\mathbf{P}\mathbf{y}$, penalising predictions on emotions opposite to the true label more strongly than predictions on unrelated emotions. We drop the clipping in the soft target:

$$\tilde{\mathbf{y}}_{\mathrm{neg}} = (1 - \alpha)\,\mathbf{y} + \alpha\,\mathbf{P}\mathbf{y}.$$

The Lovász hinge handles negative target values directly.

4.3 Curriculum on α

Training with $\alpha > 0$ from epoch one mixes hard and soft signals before the model has learned anything. We avoid this by ramping α from 0 at epoch 3 to its target value at epoch 8. Earlier epochs see only the hard target; later epochs see the full Plutchik-weighted target. This is critical: as the ablation in Section 5.3 shows, removing the curriculum causes the model to degenerate by predicting many adjacent emotions at once, which inflates the Plutchik score while collapsing F1.

4.4 Optimisation

Layer-Wise Learning Rate Decay (LLRD) The early-fusion gate and the classifier head are randomly initialised, whereas the transformer stack is pretrained, so updating them at the same learning rate degrades performance. We apply $5\times$ the base rate to the new modules and the classifier head, $1\times$ to the upper transformer layers, and $0.5\times$ to the lower layers.

Per-Class Thresholds At inference, we calibrate one threshold per emotion. For each emotion c, we sweep thresholds on the validation set and pick the value that maximises validation macro-F1.

5 Results

We present results in three parts. Section 5.1 covers the main GoEmotions experiment. Section 5.2 reports our model on the three secondary datasets. Section 5.3 ablates the components of our training apparatus across all four datasets.

5.1 GoEmotions

Table 3 compares twelve training configurations on the GoEmotions test set, each row averaged over five seeds, varying three factors: the loss function (baseline binary cross-entropy, SLS, SJL, SJL-neg), the soft-target weight α, and the base learning rate. Per-class thresholds are tuned on the validation set, following the protocol of Lecourt et al. (2026).

F1 saturates. Macro-F1 ranges from 0.5520 to 0.5603 across all twelve configurations, a total spread of less than one F1 point. Under F1, every configuration looks the same.

Plutchik discriminates. The Plutchik score ranges from 0.6690 to 0.6999, a spread of 0.0309, about 3.7 times the F1 spread. Plutchik separates configurations that F1 cannot.

Configuration	macro-F1	Prec	Rec	Plutchik
Part A — reported				
BERT-base (Demszky et al., 2020)	0.46	0.40	0.63	—
BERT-base (Kocoń et al., 2023)	0.5275	—	—	—
RoBERTa-LexEnhance (Lecourt et al., 2026)	0.5518	0.5585	0.5760	—
RoBERTa-EarlyFusion (Lecourt et al., 2026)	0.5498	0.5495	0.5779	—
Part C — RoBERTa-EarlyFusion				
baseline ($\alpha = 0$, lr $2 \cdot 10^{-5}$)	0.5551	0.5474	0.5811	0.6767
SLS, $\alpha = 0.15$, lr $5 \cdot 10^{-5}$	0.5541	0.5553	0.5732	0.6787
SLS, $\alpha = 0.15$, lr $2 \cdot 10^{-5}$	0.5598	0.5588	0.5825	0.6821
SLS, $\alpha = 0.15$, lr $1 \cdot 10^{-5}$	0.5520	0.5523	0.5767	0.6690
SJL, $\alpha = 0.15$, lr $5 \cdot 10^{-5}$	0.5534	0.5410	0.5807	0.6860
SJL, $\alpha = 0.15$, lr $2 \cdot 10^{-5}$	0.5581	0.5633	0.5757	0.6774
SJL, $\alpha = 0.15$, lr $1 \cdot 10^{-5}$	0.5547	0.5457	0.5872	0.6858
SJL-neg, $\alpha = 0.05$, lr $2 \cdot 10^{-5}$	0.5584	0.5636	0.5774	0.6841
SJL-neg, $\alpha = 0.10$, lr $2 \cdot 10^{-5}$	0.5582	0.5495	0.5850	0.6866
SJL-neg, $\alpha = 0.15$, lr $5 \cdot 10^{-5}$	0.5538	0.5389	0.5889	**0.6999**
SJL-neg, $\alpha = 0.15$, lr $2 \cdot 10^{-5}$	**0.5603**	0.5605	0.5830	0.6865
SJL-neg, $\alpha = 0.15$, lr $1 \cdot 10^{-5}$	0.5561	0.5591	0.5793	0.6727

Table 3: GoEmotions test results. Mean over 5 seeds. F1 spread is less than one point across all configurations. Plutchik spread is about 3.7 times wider. The F1-best and Plutchik-best configurations differ.

Model	J@0.5	J_{pc}	Plutchik
Part A — reported			
NTUA-SLP (Baziotis et al., 2018)	0.5788	—	—
TCS-Research (Meisheri and Dey, 2018)	0.5820	—	—
PlusEmo2Vec (Park et al., 2018)	0.5760	—	—
SpanEmo (Alhuzali and Ananiadou, 2021)	**0.601**	—	—
Part B — re-run by us			
SpanEmo (reproduction)	0.5938	—	0.6438
Part C — RoBERTa-EarlyFusion			
baseline	0.5593	0.5776	0.7904
SLS, $\alpha = 0.10$	0.5591	0.5761	0.7919
SLS, $\alpha = 0.15$	0.5572	0.5784	**0.7922**
SJL, $\alpha = 0.05$	0.5602	0.5777	0.7910
SJL, $\alpha = 0.10$	0.5610	**0.5802**	0.7906
SJL, $\alpha = 0.15$	0.5602	0.5774	0.7898
SJL-neg, $\alpha = 0.10$	0.5516	0.5676	0.7848
SJL-neg, $\alpha = 0.15$	0.5462	0.5526	0.7702

Table 4: SemEval-2018 E-c English test results. Part C: mean over 5 seeds. Part B: single seed . J_{pc} uses per-class thresholds calibrated on validation.

The two metrics disagree on the winner. The best F1 of 0.5603 comes from SJL-neg at $\alpha = 0.15$, lr $= 2 \cdot 10^{-5}$, while the best Plutchik of 0.6999 comes from the same loss and α at a different learning rate, lr $= 5 \cdot 10^{-5}$. A different learning rate gives a different winner depending on which metric you look at.

Semantic losses help on Plutchik. The four configurations with the highest Plutchik scores all use a semantic loss (SJL or SJL-neg). The baseline binary cross-entropy reaches 0.6767, while the top semantic configuration reaches 0.6999, a gain of 2.3 points.

5.2 Other Datasets

We now check whether the GoEmotions findings hold on three other emotion corpora. We compare three kinds of rows in each table: prior reported numbers (Part A); a re-run of accessible SOTA model when possible (Part B); our model (Part C).

SemEval-2018 E-c. Table 4 shows the SemEval-2018 results. The primary metric is Jaccard. Plutchik columns are empty for Part A because the Plutchik score did not exist when those papers were published; computing it for a prior model requires reproducing it to obtain per-instance predictions, which is what Part B provides for SpanEmo.

The headline finding: matched on Jaccard, far ahead on Plutchik. SpanEmo, the strongest accessible prior model, reaches $J@0.5 = 0.5938$. Our SJL configuration at $\alpha = 0.10$ reaches $J_{pc} = 0.5802$ with per-class tuning, within 0.014 Jaccard of SpanEmo. On Plutchik, the picture inverts. SpanEmo scores 0.6438. Our best configuration (SLS, $\alpha = 0.15$) reaches 0.7922, a gap of about 15 percentage points under our single-seed reproduction of SpanEmo (see Section 8). F1 and Jaccard alone would call this a loss. Plutchik does not.

Within our own configurations (Part C), the same pattern as GoEmotions emerges. SJL at $\alpha = 0.10$ achieves the best J_{pc} (0.5802). The Plutchik score is highest for the SLS configurations (0.7922, 0.7919). SJL-neg at $\alpha = 0.15$ underperforms on both metrics, consistent with the GoEmotions observation that high-α negative targets can destabilize training on multi-label sigmoid heads.

CancerEmo. CancerEmo annotates each sentence for eight emotions independently, each as a separate binary classification task. Standard practice (Sosea and Caragea, 2020, 2021; Giannouris et al., 2025) is to train one binary classifier per emotion and report the average F1 across emotions. We follow this protocol in Table 5. Plutchik does not apply here, because each model only predicts one emotion at a time; there is no multi-label prediction vector to score.

A note on the BERT reproduction. Our BERT reproduction reaches 0.7493 AVG-F1, four points above the 0.701 reported by Sosea 2020. The gap is consistent with two protocol differences. First, we calibrate per-emotion thresholds on the validation set; the original used a fixed threshold. Second, the `bert-base-uncased` weights on HuggingFace have been updated since 2020. Our reproduction therefore matches the more recent literature trajectory: BERT 0.71 (2020) → eMLM 0.72 (2021) → AnchorBERT 0.74 (2025) → our reproduction 0.75 (this paper). Within our own consistent protocol, both Part B and Part C are evaluated identically.

Model	AVG-F1	AVG-P	AVG-R
Part A — reported			
BiLSTM (Sosea and Caragea, 2020)	0.61	—	—
BERT (Sosea and Caragea, 2020)	0.701	—	—
eMLM (Sosea and Caragea, 2021)	0.718	—	—
AnchorBERT (Giannouris et al., 2025)	0.74	—	—
Part B — re-run by us			
BERT (our reproduction)	0.7493	0.6901	0.8387
eMLM (our reproduction)	0.7309	0.6557	0.8468
Part C — RoBERTa-EarlyFusion			
baseline	0.7625	0.6864	0.8749
SLS, $\alpha = 0.10$	0.7601	0.6904	0.8595
SLS, $\alpha = 0.15$	0.7612	0.6839	0.8747
SJL, $\alpha = 0.05$	0.7609	0.6895	0.8630
SJL, $\alpha = 0.10$	**0.7664**	0.6920	0.8729
SJL, $\alpha = 0.15$	0.7647	0.6931	0.8681
SJL-neg, $\alpha = 0.10$	0.7627	0.6914	0.8662
SJL-neg, $\alpha = 0.15$	0.7623	0.6869	0.8726

Table 5: CancerEmo per-emotion binary results. Part C: mean over 5 seeds. Part B: BERT 5 seeds, eMLM single seed (multi-seed deferred, see Section 8). Each row is the average across 8 independent binary classifiers, one per emotion. Plutchik is not reported because per-emotion binary evaluation does not produce a multi-label prediction vector.

	8-way		15-way	
Model	macro-F1	Plut	macro-F1	Plut
Part A — reported				
distilRoBERTa (Koufakou et al., 2024)	—	—	0.26	—
Part C — RoBERTa-EarlyFusion				
baseline	0.3292	0.2315	0.2323	0.2692
SLS, $\alpha = 0.10$	0.3295	0.2319	0.2313	0.2664
SLS, $\alpha = 0.15$	**0.3301**	**0.2331**	0.2340	0.2643
SJL, $\alpha = 0.05$	0.3292	0.2306	0.2316	0.2659
SJL, $\alpha = 0.15$	0.3299	0.2305	0.2340	0.2632
SJL-neg, $\alpha = 0.15$	0.3300	0.2311	0.2310	0.2596

Table 6: GoodNewsEveryone results. Two evaluation grains on the same model outputs: 8-way collapses native GNE labels onto the eight Plutchik primaries (the natural grain for Plutchik-grounded evaluation); 15-way keeps the native GNE label set, comparable to published GNE numbers by Koufakou et al. but mechanically lower in macro-F1 because of finer-grained class boundaries. All configurations cluster within a small band on both metrics and both grains. Single-label softmax has no near-miss gradient channel for Plutchik-aware losses to exploit.

GoodNewsEveryone. GoodNewsEveryone is a single-label news headline corpus whose label set is explicitly Plutchik-based, which is why we include it. We observe that all our configurations cluster within a 0.002 F1 band and a 0.004 Plutchik band, meaning neither metric discriminates meaningfully on this dataset. We attribute this to the single-label setup and class imbalance. In multi-label tasks, the gradient on a non-target class depends on the soft-target weight assigned to it; this is the channel through which Plutchik-aware losses act. In single-label softmax, the gradient on a non-target class depends only on the predicted probability and not on the class identity. Plutchik-aware losses have nothing to act on. We report two evaluation grains on the same model outputs. The 8-way grain collapses GNE labels onto the eight Plutchik primaries via a fixed mapping, and is the natural grain for Plutchik-grounded evaluation. The 15-way grain keeps the native GNE label set and is directly comparable to published GNE numbers. The 15-way Plutchik spread is 0.0096, more than twice the 8-way spread of 0.0042. Despite a lower F1 compared to SOTA, with more labels for the model to confuse, Plutchik has more room to register differences. Table 6 presents our results.

The Plutchik scores are also surprisingly low (~ 0.23), for structural reasons. On GoEmotions, the metric averages over many true labels per instance and the maximum-similarity match is usually high. On a single-label dataset, only one true label is available per instance, and a single wrong prediction with a non-adjacent emotion drives the score down.

5.3 Ablation Study

We ablate seven components of the training apparatus, starting from the full model (SJL-neg, $\alpha = 0.15$, $\text{lr} = 2 \cdot 10^{-5}$, with curriculum, LLRD, and per-class threshold tuning). Table 7 reports each ablation across all four datasets.

The No Curriculum collapse is GoEmotions-specific. On GoEmotions, removing the curriculum collapses F1 to 0.0794 while Plutchik shoots to 0.9275. The model has learned to predict every adjacent emotion on every example, evidence that the Plutchik score can be gamed. On SemEval, CancerEmo, and GoodNewsEveryone (both 8-way and 15-way grains), the same ablation produces results essentially identical to the full model. Curriculum non-optionality is a property of dense multi-label tasks; on smaller label spaces the failure mode does not materialise. We keep the curriculum across all datasets for protocol consistency. This is a direct consequence of using metrics that rewards partial credit and not a flaw in the score. Continuous reward signals admit specification gaming. However, we are interested in whether normal training trajectories drift toward the degenerate solution, and, with the curriculum, they do not.

Strict Threshold cost is also dataset-specific. Replacing per-class thresholds with a fixed 0.5 drops GoEmotions F1 by 6.1 points and Plutchik by 3.4 points, and CancerEmo F1 by 4.6 points. On SemEval, GoodNewsEveryone, and the GNE 15-way grain, the same change is within 0.001 on every metric. Threshold calibration matters where label imbalance is severe.

Removing Early Fusion improves Plutchik

	GoEmotions		SemEval		CancerEmo	GNE 8-way		GNE 15-way	
Ablation	F1	Plut	J@0.5	Plut	AVG-F1	F1	Plut	F1	Plut
Full Model	0.5590	0.6689	0.5602	0.7910	0.7629	0.3311	0.2318	0.2340	0.2632
No Early Fusion	0.5528	0.6888	0.5519	0.7952	0.7556	0.3296	0.2419	0.2369	0.2724
No Late Fusion	0.5608	0.6755	0.5637	0.7797	0.7600	0.3284	0.2347	0.2416	0.2919
No LLRD	0.5491	0.6712	0.5491	0.7862	0.7541	0.3272	0.2277	0.2235	0.2503
No Curriculum	**0.0794**	**0.9275**	0.5604	0.7902	0.7590	0.3314	0.2325	0.2333	0.2666
No Semantic Loss	0.5536	0.6793	0.5592	0.7920	0.7643	0.3330	0.2349	0.2323	0.2692
Strict Threshold	0.4984	0.6345	0.5610	0.7903	0.7173	0.3311	0.2318	0.2340	0.2632

Table 7: Cross-dataset ablation. Mean over 5 seeds. The catastrophic No-Curriculum collapse on GoEmotions does not appear on the smaller label spaces; the Strict-Threshold cost is concentrated on label-imbalanced GoEmotions and CancerEmo. CancerEmo Plutchik is omitted because its per-emotion binary protocol does not produce a multi-label prediction vector.

on every dataset. Dropping the General-Inquirer early-fusion gate from the architecture raises Plutchik by $+0.020$ on GoEmotions, $+0.004$ on SemEval, $+0.010$ on GNE 8-way, and $+0.009$ on GNE 15-way, at small cost in primary metric (within 0.008 F1, Jaccard, or AVG-F1). The GI features that benefit F1 marginally hurt Plutchik. This is itself evidence the Plutchik score reveals trade-offs F1 hides: even our own architectural choices are not Pareto-optimal under our own evaluation.

Other components have small, mixed effects. Removing Late Fusion, LLRD, or the Semantic Loss changes F1 and Plutchik by less than half a point in either direction across datasets. The full model is not Pareto-dominant on any single component; it is the best overall combination of small individual contributions.

6 Discussion

The Plutchik score discriminates where F1 saturates. On GoEmotions, twelve configurations span a 0.008 F1 range and a 0.031 Plutchik range. The Plutchik score is about 3.7 times more sensitive to changes in training procedure than F1 is.

F1 underestimates models that make psychologically reasonable mistakes. On GoEmotions, the F1-optimal and Plutchik-optimal configurations differ. SpanEmo and our SemEval models illustrate the broader pattern: under Jaccard the systems are within sampling noise, while under Plutchik our models score 15 percentage points higher because their errors are concentrated on emotions adjacent to the gold label rather than distributed uniformly across the wheel. F1 treats the two systems as equivalent; Plutchik makes the structural difference visible.

The Plutchik score can be gamed. Removing the curriculum during training produces a degenerate model: it maximizes Plutchik on GoEmotions (0.9275) by overpredicting adjacent emotions while F1 collapses to 0.08. This is a property of any partial-credit metric. The collapse is GoEmotions-specific (see Section 5.3); we treat the curriculum as a non-optional component for protocol consistency across datasets.

The Plutchik score requires a multi-label prediction vector. On per-emotion binary tasks (the standard CancerEmo protocol), the score does not apply. On single-label softmax tasks (GoodNewsEveryone), it applies but the model has no gradient channel through which Plutchik-aware losses can affect it. The score is most informative on multi-label tasks with sigmoid heads.

The training apparatus helps on F1 too. For readers who do not buy the score, the apparatus described in Section 4 also yields a small standalone improvement: our best GoEmotions configuration reaches 0.5603 macro-F1, a 1.0 point gain over the previously published RoBERTa baseline of 0.5498 (Lecourt et al., 2026), consistent across configurations.

7 Conclusion

We propose the Plutchik score as an additional evaluation measure for emotion classification. Across four corpora, the score discriminates between models that F1 cannot. On GoEmotions specifically, it separates configurations whose F1 differences are within noise. We show that the score can be gamed by an unstable training trajectory but that a simple curriculum prevents this. Independently of the score, the training apparatus improves over the previously published RoBERTa-EarlyFusion baseline on GoEmotions (Lecourt et al., 2026). Last but not least, please note that while the methodology will not change, other geometries inspired from psychological frameworks such as the Russel model merit future work exploration currently undertaken by the authors.

8 Limitations

The wheel geometry is hand-designed. We use Plutchik's specific arrangement, including the half-integer placement of dyads. Other geometries are possible (Russell's circumplex, the OCC model, learned embeddings). We make no claim that Plutchik's wheel is the right one. We claim only that *any* psychologically grounded geometry, applied to model evaluation, reveals structure that F1 hides. Future work should compare geometries directly.

Alternative similarity geometries we explored. Before adopting Plutchik's wheel, we experimented with two data-driven alternatives. The first defined an inter-emotion similarity matrix derived from the linguistic-signature analysis of 13 English emotion corpora reported by Lecourt et al. (2026). We did not adopt it for the present paper because the matrix is data-dependent (limited to the 13 corpora analysed, all English) and the selection of linguistic features used to define the signatures is partly arbitrary. The second was a similarity based on contextual word embeddings: for each emotion we built a representation by averaging RoBERTa embeddings of the lexicon-derived words associated with that emotion, and used cosine similarity between these representations as the inter-emotion similarity. We explored this alternative but did not adopt it as our reference geometry for this paper. A systematic comparison of these data-driven alternatives against the Plutchik geometry used here is a natural extension we leave to future work.

Reproducing prior SOTA is incomplete. We re-run three models in Part B: BERT and eMLM on CancerEmo, and SpanEmo on SemEval-2018. The eMLM and SpanEmo reproductions are single-seed because of resource constraints. NTUA-SLP for SemEval and the Koufakou model for GoodNewsEveryone are reported-only. The Plutchik score cannot be applied retroactively to reported-only models without their predictions.

English only. All four corpora are English. The Plutchik framework is psychologically universal but linguistic implementations of emotion vary across cultures. Cross-lingual evaluation is future work.

Emotion-aware annotation protocols. Our work addresses evaluation given fixed datasets and label inventories. A complementary direction, which we leave to future work, is the design of annotation protocols that capture emotion relatedness at annotation time: instead of forcing annotators into single-label or fixed multi-label decisions, soliciting graded or wheel-anchored judgements could reduce the mismatch between annotation and Plutchik-grounded evaluation.

Aggregate F1+Plutchik measure. We report F1 and Plutchik as separate columns and let the reader weigh them. An aggregate measure that combines the two into a single score (for instance, a weighted geometric mean) would allow direct ranking of systems. Designing and validating such an aggregate is future work.

9 Ethical Considerations

All four datasets we use are publicly released benchmarks; we performed no new data collection. We verified that each was anonymised at source by its original creators: GoEmotions has named entities and usernames masked (Demszky et al., 2020); SemEval-2018 distributes only tweet text with user handles stripped (Mohammad et al., 2018); CancerEmo, drawn from a cancer-survivor support forum, is released at the sentence level with user identifiers removed and no thread context that could enable re-identification (Sosea and Caragea, 2020); GoodNewsEveryone contains public news headlines with no personal data (Bostan et al., 2019). CancerEmo concerns a sensitive health context; we use it only for its intended emotion-classification benchmarking purpose, with no patient-level inference attempted.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728. Association for Computational Linguistics.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th In-*

ternational Workshop on Semantic Evaluation, pages 245–255. Association for Computational Linguistics.

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. 2018. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4413–4421. IEEE.

L. Bostan, E. Kim, and R. Klinger. 2019. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *arXiv preprint arXiv:1912.03184*.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Chih Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958. Association for Computational Linguistics.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. 2021. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15334–15342.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. abs/1810.04805.

Paul Ekman and Dacher Keltner. 1970. Universal facial expressions of emotion. *California mental health research digest*, 8(4):151–158.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: A contrast-enhanced prompt-tuning framework for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2947–2957.

Polydoros Giannouris, Vasileios Mygdalis, and Ioannis Pitas. 2025. Improving multilabel text emotion detection with emotion interrelation anchors. *Natural Language Processing Journal*, 12:100170.

Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. 2024. Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 14(21):9863.

Ziyan Hossain. 2025. Designing with grief: The pacer framework for systemic practice towards post-biological systems of care, memory, and connection. Technical report, OCAD University.

Guiyang Hou, Yongliang Shen, Wenqi Zhang, Wei Xue, and Weiming Lu. 2023. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12651–12663. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052. Association for Computational Linguistics.

Mingxuan Ju, Wei Song, Shiyu Sun, Yanfang Ye, Yujie Fan, Shifu Hou, Kenneth Loparo, and Liang Zhao. 2021. Dr.emotion: Disentangled representation learning for emotion analysis on social media to improve community resilience in the COVID-19 era and beyond. In *Proceedings of the Web Conference 2021*, pages 518–528. ACM.

J. Kocoń and 1 others. 2023. ChatGPT: Jack of all trades, master of none. 99:101861.

Anna Koufakou, Elijah Nieves, and John Peller. 2024. Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 196–206. Association for Computational Linguistics.

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. 'Only ChatGPT gets me': An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2603–2611. ACM.

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2026. Linguistic Signatures for Enhanced Emotion Detection.

Dongjun Lim and Yun-Gyung Cheong. 2024. Integrating plutchik's theory with mixture of experts for enhancing emotion classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 857–867. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Yuzhao Mao, Di Lu, Xiaojie Wang, and Yang Zhang. 2023. FATRER: Full-attention topic regularizer for accurate and robust conversational emotion recognition. *Preprint*, arxiv:2307.12221 [cs].

Manish Marwah, Asad Narayanan, Stephan Jou, Martin Arlitt, and Maria Pospelova. 2024. Is f_1 score suboptimal for cybersecurity models? introducing $c_\{score\}$, a cost-aware alternative for model assessment. *arXiv preprint arXiv:2407.14664*.

Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* the MIT Press.

Hardik Meisheri and Lipika Dey. 2018. Tcs research at semeval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th international workshop on semantic evaluation*, pages 291–299.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *SemEval*, pages 1–17, New Orleans, Louisiana. ACL.

Matthew Morris, David J Tena Cucala, Bernardo Cuenca Grau, and Ian Horrocks. 2024. Relational graph convolutional networks do not learn sound rules. *arXiv preprint arXiv:2408.10261*.

Donovan Ong, Shuo Sun, Jian Su, and Bin Chen. 2024. Mitigating linguistic artifacts in emotion recognition for conversations from tv scripts to daily conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11319–11324.

Ji Ho Park, Peng Xu, and Pascale Fung. 2018. Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and# hashtags. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 264–272.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions.

Robert Plutchik. 1991. *The emotions*. Bloomsbury Publishing PLC.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Yu-Ping Ruan and Taihao Li. 2023. Capsule network with label dependency modeling for multi-label emotion classification. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Rădulescu, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

N. Setyoningrum, N. SM, A. Alam, A. Nurdin, D. Nursamsi, and M. Lodana. 2025. The impact of linguistic features on emotion detection in social media texts. *Innovation in Research of Informatics (Innovatics)*.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560. Association for Computational Linguistics.

T. Sosea and C. Caragea. 2020. Canceremo: A dataset for fine-grained emotion detection. In *EMNLP*, pages 8892–8904.

Tiberiu Sosea and Cornelia Caragea. 2021. eMLM: A new pre-training objective for emotion related tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 286–293. Association for Computational Linguistics.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Intensity-valued emotions help stance detection of climate change twitter data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6246–6254. International Joint Conferences on Artificial Intelligence Organization.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722.

Wen Wu, Bo Li, Chao Zhang, Chung-Cheng Chiu, Qiujia Li, Junwen Bai, Tara Sainath, and Phil Woodland. 2024. Handling ambiguity in emotion: From out-of-domain detection to distribution estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2093. Association for Computational Linguistics.

Lizhi Xiong, Jianhua Xu, Ching-Nung Yang, and Xinpeng Zhang. 2023. Cmcf-net: An end-to-end context multiscale cross-fusion network for robust copy-move forgery detection. *IEEE Transactions on Multimedia*, 26:6090–6101.

Bo Xu, Longjiao Li, Wei Luo, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2024. Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2318–2328.

Yige Xu, Zhiwei Zeng, and Zhiqi Shen. 2023. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11654–11666.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421. Macao.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023a. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408.

Ting Zhang, Zhuang Chen, Ming Zhong, and Tieyun Qian. 2023b. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. *Preprint*, arxiv:2306.06601 [cs].

Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022. Mucdn: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582. Association for Computational Linguistics.