scientific reports

Check for updates

OPEN

A weighted difference loss approach for enhancing multi-label classification

Qiong Hu[1,2]✉, Masrah Azrifah Azmi Murad[1,3], Azreen Bin Azman[1,3] & Nurul Amelina Nasharuddin[1,3]

Conventional multi-label classification methods often fail to capture the dynamic relationships and relative intensity shifts between labels, treating them as independent entities. This limitation is particularly detrimental in tasks like sentiment analysis where emotions co-occur in nuanced proportions. To address this, we introduce a novel Weighted Difference Loss (WDL) framework. WDL operates on three core principles: (1) transforming labels into a normalized distribution to model their relative proportions; (2) computing learnable, weighted differences across this distribution to explicitly capture inter-label dynamics and trends; and (3) employing a label-shuffling augmentation to ensure the model learns intrinsic, order-invariant relationships. Our framework not only achieves state-of-the-art performance on four public benchmarks, but more importantly, it substantially improves the recognition of minority classes. This demonstrates the framework's ability to learn from sparse data by effectively leveraging the underlying label structure, offering a robust, loss-driven alternative to complex architectural modifications.

Keywords Multi-label Sentiment Classification, BERT, Loss Function Optimization, Label Dependency Modeling, Weighted Difference Loss (WDL)

Multi-label sentiment classification (MLSC) is a critical task for understanding the nuanced and often complex emotions expressed in text, with applications ranging from market research to public opinion analysis[1–4]. Unlike single-label tasks, MLSC acknowledges that a single text can convey multiple sentiments simultaneously[5]. However, this task is impeded by two persistent challenges: severe class imbalance, where minority emotions are poorly learned[6], and the flawed label independence assumption inherent in standard fine-tuning approaches. Pre-trained models like BERT[7], despite their power, often inherit this limitation by using loss functions like Binary Cross-Entropy (BCE), which by design treats each label as a separate binary problem, thus failing to model the rich interdependencies between them[5,8].

This failure to model label relationships is not merely a statistical issue; it represents a fundamental misunderstanding of sentiment. Emotions are not independent events but exist in a structured, dynamic relationship[9]. For instance, an increase in 'joy' often corresponds to a decrease in 'sadness'[10], and the co-occurrence of 'joy' and 'surprise' has a different proportional intensity than 'anger' and 'disgust'. To overcome these limitations, we argue for a paradigm shift: from predicting independent label probabilities to modeling a structured label distribution[11]. Our core hypothesis is that by supervising not only the presence of labels but also their relative proportions and their rates of change (differences), a model can learn the underlying structure of the label space without requiring external knowledge graphs or complex architectural changes[12,13].

To operationalize this paradigm shift, we propose the Weighted Difference Loss (WDL) framework. This paper makes the following primary contributions:

- We introduce a novel ratio-to-difference mechanism that normalizes label values into a distribution of relative proportions and then computes higher-order differences to explicitly model the dynamic trends and interdependencies between labels.
- We design a learnable weighting scheme that allows the model to adaptively balance the supervisory signals from the base classification loss, the ratio-matching loss, and the multi-order difference losses, thereby optimizing the learning process.

[1]Faculty of Computer Science and Information Technology, UPM Lebuh Universiti, 43400 Serdang, Selangor, Malaysia. [2]Communication and Computer Science College, Nanjing Tech University Pujiang Institute, Lixue Road, 211200 Nanjing, Jiangsu, China. [3]Masrah Azrifah Azmi Murad, Azreen Bin Azman and Nurul Amelina Nasharuddin: These authors contributed equally to this work. ✉email: gs65254@student.upm.edu.my

- We incorporate a label-shuffling augmentation strategy during training, which forces the model to learn intrinsic, order-invariant relationships between emotions, significantly enhancing its robustness and generalization capabilities.
- We empirically demonstrate through extensive experiments on four public benchmarks that our WDL framework achieves state-of-the-art performance and, most critically, substantially improves the recognition of minority classes, evidenced by a 0.90 absolute F1-score gain for the 'grief' category on the GoEmotions dataset.

The remainder of this paper is organized as follows: Section "Related work" reviews related work. Section "The WDL framework" details the proposed WDL framework. Section "Experimental analysis" presents the experimental setup and results. Section "Ablation study and discussion" provides ablation studies and discussion. Finally, Section "Conclusion" concludes the paper.

Related work

Sentiment analysis aims to automatically extract subjective information from text[14,15]. Multi-label sentiment classification (MLSC), a subfield, addresses the realistic scenario where a text expresses multiple, intertwined emotions[16,17]. The evolution of MLSC methods reflects a continuous effort to better capture textual context and label relationships.

Evolution and persistent challenges in MLSC

Early approaches relied on traditional feature engineering (e.g., N-grams, TF-IDF), which required significant manual effort and lacked deep contextual understanding[18,19]. The advent of deep learning models like CNNs and RNNs automated feature extraction but often struggled with long-range dependencies and implicitly assumed label independence[20,21].

The introduction of Transformer-based Pre-trained Language Models (PLMs), particularly BERT[7], revolutionized the field with powerful contextual representations[22]. However, even when fine-tuned, PLMs still face two core MLSC challenges: (1) Class Imbalance, where models become biased towards frequent emotions[23], and (2) the Label Independence Assumption, where standard loss functions like BCE neglect the rich, natural correlations between emotions[24].

Modern strategies for enhanced MLSC

Contemporary research has explored various strategies to overcome these limitations, as summarized in Table 1. Our work primarily contributes to the "Loss Function Modification" category, but its prompt-based input formulation also connects it to "Advanced Representation" techniques.

Innovations in loss functions directly steer model training. Focal Loss[25] and ASL[26] address class imbalance by re-weighting examples. LDL[27] learns a probability distribution over labels, implicitly modeling relationships. While effective, these methods may not fully capture the relative proportional strength or dynamic shifts between co-occurring emotions.

Explicit modeling of label dependencies directly represents label relationships. GNNs[31] are a dominant paradigm, constructing a label graph (from co-occurrence statistics or external knowledge) and propagating information to learn correlation-aware predictions. While powerful, GNN-based methods introduce significant overheads: they require the pre-construction of a label graph, which may be suboptimal or unavailable, and add notable computational complexity[35]. To circumvent these issues, we propose an alternative, loss-driven approach. Instead of encoding label relationships into a fixed graph structure, our method forces the model to learn these relationships dynamically from the data itself, guided purely by the loss function.

Advanced representation learning techniques, such as contrastive learning[33] and prompt-based learning[34], aim to improve the underlying features. Contrastive methods learn more discriminative embeddings by pushing dissimilar samples apart in the feature space. Prompting reformulates the task to better align with the PLM's pre-training objectives. Our work incorporates a prompt-inspired input formulation but focuses its core innovation on the loss function, making it complementary to these representation-focused methods.

Motivation for weighted difference loss

Existing approaches often specialize in either class imbalance or label dependency, introduce significant architectural complexity, or fail to model the nuanced, relative proportional strengths of emotions. Our proposed WDL offers a unified, lightweight solution that operates directly on the model's output distribution. By focusing on the learnable, weighted differences in normalized label proportions, WDL provides a computationally tractable method to simultaneously mitigate class imbalance and model label interdependencies, aiming to improve performance, particularly for minority classes.

Approach Category	Representative Methods / Key Papers	Core Strategy Keywords for Imbalance / Dependencies
Loss Function Modification	Focal Loss[25], Asymmetric Loss (ASL)[26], Label Distribution Learning (LDL)[27], Label Correlation Losses[28]	Re-weighting, Decoupling, Distribution Learning, Direct Correlation Terms.
Explicit Dependency Modeling	CRFs[29], Label Embeddings (LEAM[30]), Graph Neural Networks (GNNs) (Structure-based[31], KG-enhanced[32])	Graphical Models, Semantic Embeddings, Graph Propagation, Knowledge Infusion.
Advanced Representation	Contrastive Learning (Label-aware[33]), Prompt-based Learning (PLM/LLM Prompts[34])	Discriminative Features, Text-Label Alignment, Task Reformulation, In-context Learning.

Table 1. High-level overview of modern approaches in multi-label sentiment classification.

The WDL framework

Framework overview

The WDL framework enhances a standard BERT model by introducing a multi-component loss function that supervises the model on label presence, relative proportions, and inter-label trends. Figure 1 illustrates the overall workflow. Given an input text and a set of labels, the framework proceeds in three steps: 1. Prediction: A prompt-based BERT model with a feature refinement module generates predicted logits $\hat{y}_l$ for each label. 2. Transformation: Both predicted logits and true labels y are transformed into normalized ratio vectors, $\hat{r}$ and r, respectively. Higher-order differences ($\Delta^d\hat{r}$, $\Delta^d r$) are then computed from these ratio vectors. 3. Weighted Loss Calculation: A final loss, $\mathcal{L}_{\text{WDL}}$, is computed as a dynamically weighted sum of the binary classification loss, the ratio-matching loss, and the difference losses.

The complete loss function is defined as:

$$\mathcal{L}_{\text{WDL}} = w_l \cdot \text{BCE}(\hat{y}_l, y_{\text{bin}}) + \sum_{d=0}^{D} w_d \cdot \text{MSE}(\Delta^d\hat{r}, \Delta^d r) \tag{1}$$

where $w = [w_l, w_0, \ldots, w_D]$ are learnable weights, $\hat{y}_l$ are the predicted logits, y_{bin} are the true binary labels, and $\Delta^d\hat{r}$ and $\Delta^d r$ are the d-th order differences of the predicted and true ratios, respectively.

Prompt-based input and feature refinement

Inspired by prompt-based learning[36], we construct inputs by prepending emotion labels with `[MASK]` tokens to the text: "e_1 `[MASK]` ... e_M `[MASK]` . t_i". After extracting the final-layer representations for each `[MASK]` token from BERT, we hypothesize that these initial representations can be further refined to be more discriminative. To this end, we employ a Self-Attention Network (SAN) module to act as a feature refiner. Each `[MASK]` token's representation is independently processed by the SAN to enhance its contextual features before being passed to the classifiers. As our ablation study confirms (Section Component effectiveness analysis), this refinement step creates a higher-quality substrate for the WDL and is crucial for overall performance.

Ratio and difference formulation

From a theoretical standpoint, we posit that the set of co-occurring emotions in a text can be viewed as a discrete signal over the label space. The value at each point corresponds to the intensity of an emotion. The first-order difference (Δ^1) of this signal approximates its derivative-the rate of change in intensity from one emotion to the next. The second-order difference (Δ^2) approximates the second derivative, or the "acceleration" of this change. By supervising these derivatives, we compel the model to learn not just the static presence of emotions (the "position" of the signal), but also their dynamic relationships and trends (the "velocity" and "acceleration").

To operationalize this, we first transform both true labels and predicted logits into ratio vectors. For a given instance with a multi-hot true label vector $y_{\text{bin}} \in \{0, 1\}^M$, the true ratio vector r is computed by L1 normalization:

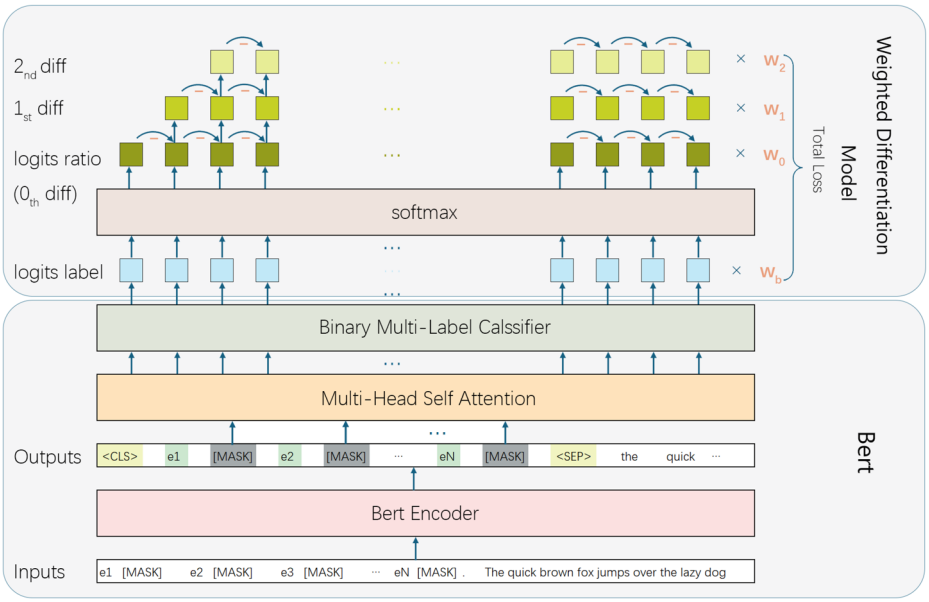

Fig. 1. The Bert-WDL architecture. An input text is prepended with emotion-guided [MASK] prompts. BERT generates representations, which are refined by a Self-Attention Network (SAN) module. The final logits are supervised by the multi-component WDL, which includes losses on labels, ratios, and their differences (2nd-order shown).

$$\mathrm{r} = \frac{\mathrm{y}_{\mathrm{bin}}}{||\mathrm{y}_{\mathrm{bin}}||_1} \tag{2}$$

If an instance has no positive labels ($||\mathrm{y}_{\mathrm{bin}}||_1 = 0$), r is a zero vector. The predicted logits $\hat{y}_l$ are passed through a softmax activation to produce a probability distribution $\hat{\mathrm{r}}$, ensuring it is also L1-normalized.

To explicitly model label dependencies, we then compute the d-th order forward difference $\Delta^d \mathrm{r}$ recursively:

$$\Delta^d \mathrm{r}[i] = \Delta^{d-1}\mathrm{r}[i+1] - \Delta^{d-1}\mathrm{r}[i] \tag{3}$$

where $\Delta^0 \mathrm{r} \triangleq \mathrm{r}$. The 1st-order difference captures intensity transitions between adjacent labels, while higher orders encode more complex, non-local dependencies.

Order-invariant learning via label shuffling

Since the difference calculation is sensitive to label order, we introduce a crucial augmentation step. During each training iteration, the original batch is expanded by creating K random permutations of the emotion label sequence for each sample. The input prompts, true labels, and true ratios are re-constructed according to these permutations, forming an augmented batch of size $K \times \mathrm{batch\_size}$. The WDL loss is computed over this entire augmented batch in a single forward and backward pass. This procedure forces the model to learn true semantic correlations between emotions (e.g., 'joy' and 'excitement') rather than spurious positional artifacts (e.g., 'the 5th label is always higher than the 4th'), thereby improving model robustness and generalization. During inference, predictions from shuffled sequences are re-ordered back to their original label sequence before evaluation.

Learnable multi-component loss

The final WDL (Eq. 1) combines the losses from the binary classification task (BCE) and $D+1$ orders of ratio/difference matching (MSE). The weights w are not fixed hyperparameters but are learned dynamically. They are parameterized by a vector of logits $\mathrm{u} \in \mathbb{R}^{D+2}$, such that $\mathrm{w} = \mathrm{softmax}(\mathrm{u})$. Both the model parameters θ and the weight logits u are updated via gradient descent, allowing the framework to adaptively determine the importance of each loss component. The complete training process is detailed in Algorithm 1.

Require: Pre-trained BERT, difference order D, permutations K, learning rates α_θ, α_w, patience P
Initialize model parameters θ, weight logits $\mathbf{u} \in \mathbb{R}^{D+2}$, best val loss $\mathcal{L}^* \leftarrow \infty$, wait $p \leftarrow 0$
for epoch $= 1, \ldots, E_{\max}$ **do**
  for each batch B **do**
    ▷ Augment batch via label shuffling
    $B_{aug} \leftarrow \emptyset$
    for each sample $s_i \in B$ **do**
      for $k = 1, \ldots, K$ **do**
        Generate a random permutation P_k of labels
        Create augmented sample s'_{ik} based on P_k
        $B_{aug} \leftarrow B_{aug} \cup \{s'_{ik}\}$
      end for
    end for
    ▷ Single forward pass on augmented batch
    Compute predicted logits $\hat{\mathbf{y}}_l = f_\theta(B_{aug})$
    Normalize predictions and targets to ratios $(\hat{\mathbf{r}}, \mathbf{r})$, then compute differences $(\Delta^d\hat{\mathbf{r}}, \Delta^d\mathbf{r})$
    Compute $\mathcal{L}_{\mathrm{WDL}}$ using Equation 1 with weights $\mathbf{w} = \mathrm{softmax}(\mathbf{u})$
    Update model parameters: $\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \mathcal{L}_{\mathrm{WDL}}$
    Update weight logits: $\mathbf{u} \leftarrow \mathbf{u} - \alpha_w \nabla_u \mathcal{L}_{\mathrm{WDL}}$
  end for
  Evaluate validation loss $\mathcal{L}_{\mathrm{val}}$
  if $\mathcal{L}_{\mathrm{val}} < \mathcal{L}^*$ **then**
    Save best θ, $\mathcal{L}^* \leftarrow \mathcal{L}_{\mathrm{val}}$, $p \leftarrow 0$
  else
    $p \leftarrow p + 1$;
    if $p \geq P$ **thenbreak**
    end if
  end if
end for

Algorithm 1. Training the WDL framework.

Implementation details

Our framework was implemented in PyTorch 2.0.0 and run on an Ubuntu 20.04 system with a 48GB vGPU. We fine-tuned *bert-base-uncased* and *bert-base-chinese* models from Hugging Face. The architecture includes a single-layer SAN for feature refinement and employs a dedicated SGD optimizer for the loss weight logits. All experiments were conducted using the comprehensive set of hyperparameters detailed in Table 2, with early stopping based on validation loss to prevent overfitting.

Experimental analysis

Datasets

We evaluated our method on four public multi-label emotion datasets: two Chinese (NLPCC 2018 Task 1 with 5 emotion labels and Ren-CECPs with 8 labels) and two English (GoEmotions with 28 labels and SemEval 2018 Task 1, E-c with 11 labels). For NLPCC, GoEmotions, and SemEval, we used the official train/validation/test splits. For datasets lacking official splits, such as Ren-CECPs, we randomly partitioned the data into training (70%), validation (15%), and test (15%) sets. To ensure reproducibility, all random partitioning was performed using a fixed random seed (42).

Evaluation metrics

We use a comprehensive suite of metrics: Macro-F1 (MF1) and Micro-F1 (mF1) to assess classification performance, with MF1 being particularly sensitive to minority class performance. We also report Average Precision (AP), Hamming Loss (HL), Coverage Error (CE), and Ranking Loss (RL). Arrows (↑ / ↓) indicate the desired direction for each metric.

Baseline methods

We compare Bert-WDL against state-of-the-art models including prompt-based (PC-MTED[37]), capsule network (CapsLDM[38]), neural architecture (MEDA-FS[39], LEM[40], EduEmo[41]), and hybrid methods (Hybrid HEF-DLF[42], Seq2Emo[43]). All baseline results are sourced from their original publications. In the following tables, a dash (-) indicates that a specific metric was not provided in the source paper.

Experimental results

Cross-dataset performance

Table 3 and Table 4 show that the WDL framework consistently delivers top-tier performance across all four datasets, demonstrating its robustness and generalizability. Unlike baseline methods that excel on one dataset but falter on another, WDL variants consistently rank among the top performers. For example, WDL2 achieves the best MF1 and mF1 on NLPCC, while WDL1 is highly competitive on Ren-CECPs and SemEval, and secures the best MF1 and mF1 on GoEmotions. This stability highlights the effectiveness of modeling label dynamics as a general principle.

Effectiveness on minority classes

The primary strength of WDL lies in its ability to mitigate class imbalance. The heatmap in Fig. 2 provides a clear visual proof of this effect on the 28-category GoEmotions dataset. In the figure, emotions are sorted by their training sample count, from the least frequent at the top to the most frequent at the bottom. This arrangement vividly illustrates that the most significant performance gains, indicated in green, occur on minority classes.

The exceptional performance on 'grief' (F1-score of 0.91 vs. a baseline of 0.01), despite only 6 training samples, strongly validates our core hypothesis. A standard BCE loss struggles with such extreme sparsity. However, WDL forces the model to consider 'grief' in relation to other emotions. By learning the difference patterns-how the presence of 'grief' alters the proportions of 'sadness' or 'disappointment'-the model can effectively infer its presence even from minimal direct evidence. This pattern of significant gains is consistent across most low-to-mid frequency emotions. While some high-frequency emotions like 'gratitude' and 'remorse' show a trade-off, indicated in red, the overall 17.4% improvement in MF1 (0.46-0.54) confirms a more balanced and robust predictive capability across the entire emotion spectrum. This is further detailed in Table 5.

Comparison of loss functions

To isolate the effect of our loss design, we compared WDL1 against standard multi-label loss functions on GoEmotions, keeping the model architecture fixed. As shown in Table 6 and the conceptual gain plot in Fig. 3, WDL1 consistently outperforms BCE, ASL, and Focal Loss in terms of both MF1 and mF1. While ASL achieves higher recall and Focal Loss higher precision, WDL1 provides the best balance, validating that explicitly modeling label dynamics is more effective than only re-weighting for class imbalance. Wasserstein loss performed poorly, suggesting it is ill-suited for this classification task without significant tuning.

Computational cost analysis

To assess the practical viability of our framework, we analyze its computational cost relative to a standard BERT baseline on the GoEmotions dataset (Table 7). Our Bert-WDL model introduces a modest increase in parameters (from 110M to 112.4M) due to the SAN module. The primary overhead comes from the label shuffling strategy ($K = 3$), which triples the number of forward passes per batch. This results in a reduction in training throughput (from 158.4 to 53.1 samples/sec) and a corresponding increase in training time per epoch. However, this is a direct and worthwhile trade-off for the substantial gains in minority class recognition and overall robustness. In contrast, the inference cost remains comparable to a standard BERT model, as shuffling is not required during evaluation. The theoretical complexity is dominated by the Transformer's $O(NL^2D)$, with the WDL component adding a negligible $O(NKD_{diff})$ term.

Parameter	Value
General Hyperparameters	
Batch size	16
Patience (P)	5
Model learning rate (α_θ)	3×10^{-5}
Weight learning rate (α_w)	0.01
Maximum sequence length	512
Maximum training epochs	100
Warmup proportion	0.1
Weight decay	0.01
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	1×10^{-8}
Random seed	42
WDL Framework Hyperparameters	
Label permutations (K)	3
Max difference order (D)	Varied in {0, 1, 2, 3}
Weight optimizer	SGD (momentum=0.9)
SAN Module Hyperparameters	
Hidden dimension (d_{model})	768
Attention heads	12
Dropout rate	0.1

Table 2. Training hyperparameters.

	Baseline Methods					Bert-WDL Variants			
Dataset	**Metric**	[39]	[37]	[40]	[41]	**WDL0**	**WDL1**	**WDL2**	**WDL3**
NLPCC	mF1 (% ↑)	63.32	64.01	-	-	66.30	66.83	**67.39**	65.70
	MF1 (% ↑)	49.23	52.69	-	-	62.67	63.12	**63.32**	62.34
	AP (% ↑)	77.19	**87.55**	-	-	73.06	72.97	73.19	62.05
	HL (↓)	0.18	-	-	-	0.13	0.13	**0.12**	0.13
	CE (↓)	1.73	1.23	-	-	**0.73**	0.74	0.75	0.78
Ren-CECPs	mF1 (% ↑)	60.76	**66.27**	50.10	62.40	62.79	62.93	62.73	62.95
	MF1 (% ↑)	48.31	54.32	44.80	56.60	**57.20**	57.11	57.04	56.69
	AP (% ↑)	76.51	**82.33**	75.10	-	66.00	66.16	66.07	65.96
	HL (↓)	**0.12**	-	0.15	-	0.14	0.14	0.13	0.13
	CE (↓)	2.22	1.89	-	-	1.41	**1.40**	**1.40**	1.42
Semeval	mF1 (% ↑)	-	-	67.50	**71.70**	71.07	71.11	70.24	70.77
	MF1 (% ↑)	-	-	56.70	58.80	57.17	**59.12**	58.25	56.40
	Jaccard (% ↑)	-	-	-	60.60	60.51	**60.66**	59.84	60.35

Table 3. Performance comparison across various difference orders.

Ablation study and discussion

Component effectiveness analysis

We conducted extensive ablation studies on the GoEmotions dataset to dissect the WDL framework and validate the contribution of each component. The results, detailed in Table 8, systematically compare variants by removing or altering key elements: the SAN for feature refinement, the learnable weights (WDL vs. D series), and the difference order. The base 'Bert' model (BERT-base with a simple classifier) serves as the fundamental baseline.

The results reveal a clear synergistic effect. First, comparing the learnable weight models (e.g., 'WDL1') against their unweighted counterparts ('D1') shows that the adaptive weighting is critical. 'WDL1' (MF1=52.18%) outperforms 'D1' (MF1=50.27%) by 1.91 absolute points, demonstrating that allowing the model to balance loss components is superior to a fixed combination.

Second, the SAN module for feature refinement provides a significant boost. 'SAN + WDL1' (MF1=53.55%) outperforms 'WDL1' without the SAN (MF1=52.18%) by 1.37 absolute MF1 points. This supports our

	Baseline Methods				Bert-WDL1
Dataset	**Metric**	42	38	43	
Semeval	mF1 (% ↑)	**78.55**	70.7	70.02	71.11
	MF1 (% ↑)	**65.77**	59.4	51.91	59.12
	Jaccard (% ↑)	**68.4**	-	-	60.66
GoEmotions	mF1 (% ↑)	-	59.3	59.57	**60.58**
	MF1 (% ↑)	49	52.7	47.28	**53.55**
	Precision (% ↑)	**54**	-	-	53.57
	Recall (% ↑)	-	-	-	**58.07**
	Jaccard (% ↑)	53.45	-	-	**57.45**

Table 4. Performance comparison with 1st-order difference model.

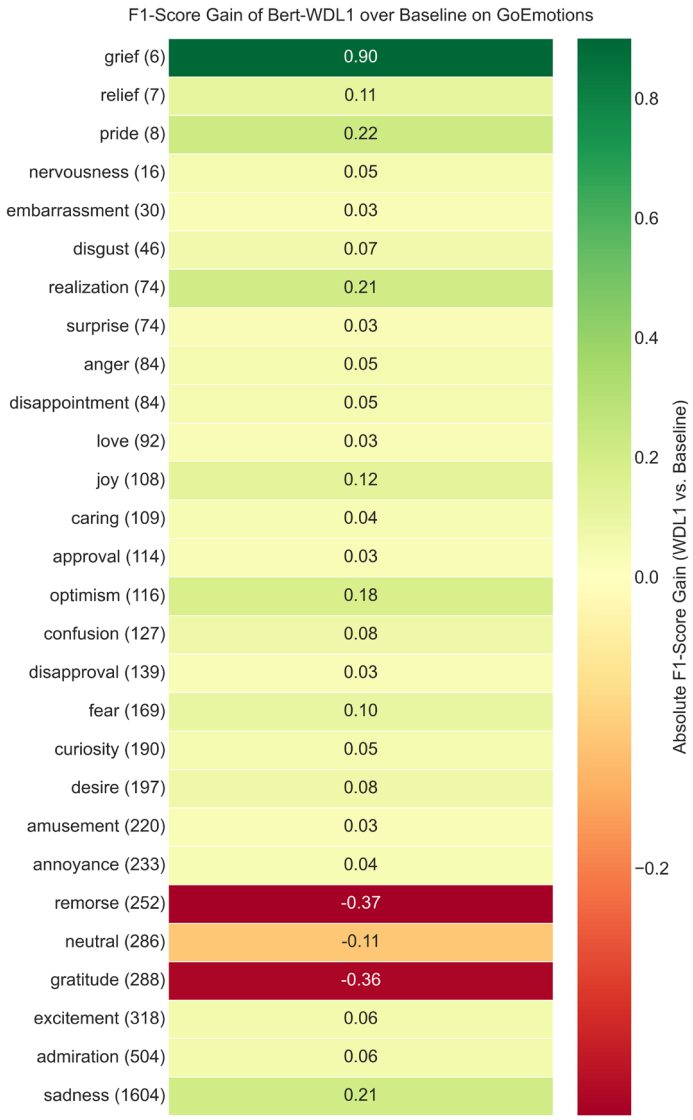

Fig. 2. Heatmap illustrating the F1-score gain of Bert-WDL1 over the baseline on the GoEmotions dataset. Emotions are sorted by their training sample count (from lowest to highest) to visualize the strong performance gains on minority classes (green) and the trade-offs on some majority classes (red).

Train Count	Emotion	Baseline[44]	Bert-WDL1
6	grief	0.01	**0.91**
7	relief	0.15	**0.26**
8	pride	0.36	**0.58**
504	admiration	0.65	**0.71**
288	gratitude	**0.86**	0.50
252	remorse	**0.66**	0.29
1604	sadness	0.49	**0.70**

Table 5. Performance comparison of Bert-WDL1 and baseline on different emotions (GoEmotions Dataset). The table showcases results for a selection of emotions, focusing on the least frequent categories to highlight improvements on minority classes.

Loss Function	MF1 (%) ↑	mF1 (%) ↑	Precision (%) ↑	Recall (%) ↑
BCE	51.59	59.93	49.73	56.57
ASL	52.31	58.42	50.04	**70.19**
Focal	51.38	58.03	**57.63**	58.43
Wasserstein	7.56	8.07	4.20	1.00
WDL1	**53.55**	**60.58**	53.57	58.07

Table 6. Performance comparison of different loss functions on emotion recognition (GoEmotions dataset).

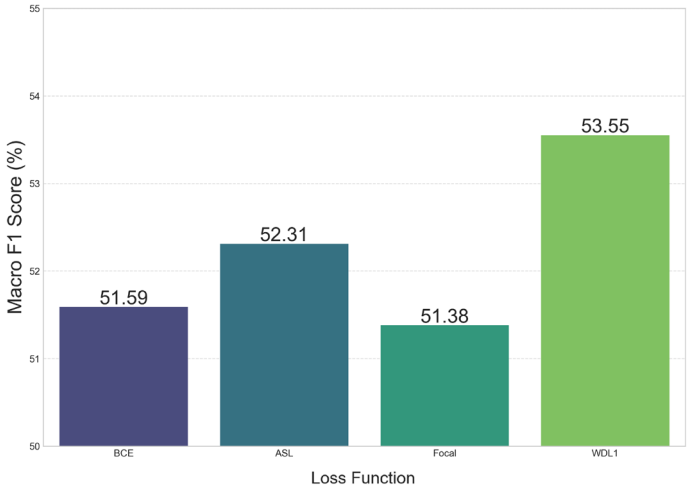

Fig. 3. Bar chart showing the relative percentage improvement in Macro-F1 score of WDL1 compared to other loss functions (BCE, ASL, Focal) on the GoEmotions dataset.

hypothesis that the SAN creates richer, more discriminative emotion representations, which in turn provides a higher-quality substrate for the WDL to operate on. Without well-defined features, calculating differences might be noisy; the SAN sharpens these features, allowing the difference loss to capture meaningful trends more effectively. The full model ('SAN + WDL1') achieves the best Macro F1, showcasing the importance of both feature refinement and learnable difference loss.

Impact of backbone model scale

To assess the scalability of our WDL framework and understand its interaction with more powerful encoders, we conducted an additional set of experiments replacing the *bert-base-uncased* backbone with its larger counterpart, *bert-large-uncased*. The results, presented in Fig. 4, reveal a nuanced relationship between model scale and performance, rather than a simple monotonic improvement.

As shown in Fig. 4, employing BERT-Large can lead to a higher peak performance. For instance, on the SemEval dataset, the BERT-Large model achieves a significantly higher peak MF1 score (approx. 59.1%) compared to the relatively flat performance of the BERT-Base model. This suggests that a larger model has the capacity to better leverage the WDL framework to capture more complex label dynamics under certain configurations.

Model	Parameters	Train Samples/sec	Inference ms/batch
BERT-base + BCE	110 M	158.4	101.0
Bert-WDL	112.4 M	53.1	105.2

Table 7. Computational cost analysis.

Model	MF1 (%) ↑	mF1 (%) ↑	Precision (%) ↑	Recall (%) ↑
SAN + WDL3	51.56	60.63	54.14	**65.51**
SAN + WDL2	52.13	60.63	**58.18**	53.32
SAN + WDL1	**53.55**	60.58	53.57	58.07
SAN + WDL0	53.08	59.96	51.14	60.16
WDL2	51.68	**61.58**	57.05	51.33
WDL1	52.18	60.98	52.30	55.12
WDL0	51.80	60.85	50.69	55.65
D2	48.37	59.54	48.00	53.64
D1	50.27	59.50	48.84	55.78
D0 (label + ratio)	51.59	59.93	49.73	56.57
Bert	44.81	59.22	56.46	39.84

Table 8. Ablation study on GoEmotions dataset.

However, the performance gains are not consistent. On the SemEval mF1 metric, the BERT-Base model consistently outperforms BERT-Large in three out of four configurations. Similarly, on the GoEmotions MF1 metric, the performance of BERT-Large is more volatile and is surpassed by BERT-Base at one of the configuration points. This indicates that simply increasing the model size does not guarantee superior performance and may even introduce instability, possibly due to overfitting or a more challenging optimization landscape.

This analysis underscores an important trade-off: while a larger backbone offers the potential for higher peak performance, it comes at a significant computational cost and without a guarantee of consistent improvement across all metrics and datasets. The choice of backbone model should therefore be considered in the context of the specific application's requirements for both performance and efficiency. This finding suggests that the primary benefits observed in our study stem from the WDL framework itself, which proves effective on both base and large model scales, rather than from simply using a larger model.

Weight dynamics and order effects

Figure 5 visualizes the learned weight distributions, revealing two key patterns. First, the weight for the 'label' component remains remarkably stable across all configurations, acting as a prediction anchor. Second, our weight parameterization scheme is designed to impose a structural prior where weights for higher-order differences decay monotonically. This design choice reflects the hypothesis that lower-order differences (e.g., 'd1') contain the most valuable signal for capturing label dynamics, while complex, higher-order interactions are progressively down-weighted to prevent the amplification of noise. As Fig. 5 confirms, the first-order difference ('d1') in the WDL1 model consequently receives a significant weight, which correlates with its strong performance on several benchmarks.

Analysis of performance trade-offs and limitations

Despite its strong performance, particularly on minority classes, our analysis reveals an important performance trade-off. As seen in Table 5, while WDL significantly boosts F1 scores for rare emotions like 'grief', it can lead to a performance decrease for some high-frequency, semantically distinct emotions like 'gratitude' and 'remorse'. We posit that this is a consequence of WDL's implicit attention re-allocation. By forcing the model to learn the relationships and relative proportions across all labels, WDL effectively redistributes the model's capacity from "over-learned" majority classes to under-represented minority classes. This is beneficial for overall balanced accuracy (MF1) but can come at the cost of peak performance on specific, well-represented labels. This trade-off highlights a key challenge for future work: developing more dynamic weighting schemes that can adapt at an instance level.

Furthermore, our experiments indicate a performance plateau or even degradation with higher-order differences ($D > 2$). We hypothesize this is due to two factors: 1) a noise amplification effect, where higher-order derivatives become overly sensitive to small perturbations in the predicted ratios, and 2) semantic sparsity, where meaningful third-order or higher emotional dependencies are rare in natural language and thus difficult to learn from limited data. This suggests that simply increasing the order is not a viable path for improvement. Future research could explore adaptive order selection mechanisms or apply regularization techniques to stabilize the learning of higher-order differences.

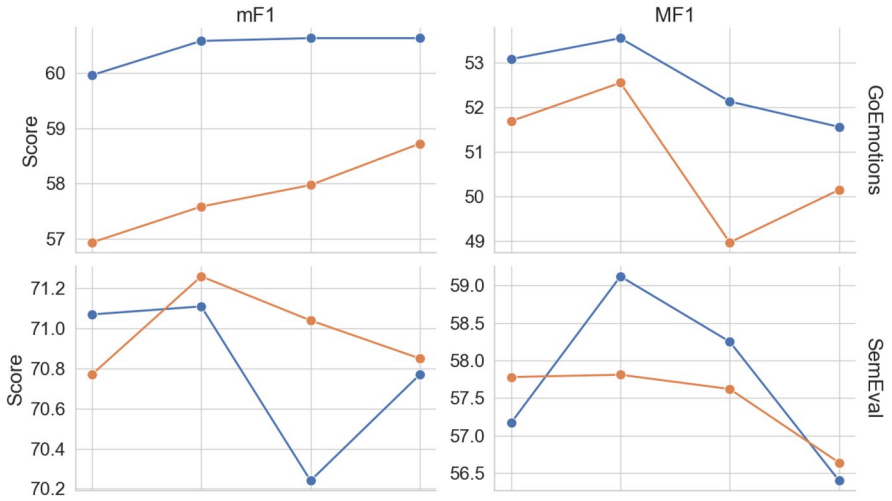

Fig. 4. Performance comparison of Bert-WDL using BERT-Base (orange line) and BERT-Large (blue line) backbones. The plots show mF1 and MF1 scores across different WDL difference orders (0 to 3 on the x-axis) for the GoEmotions (top row) and SemEval (bottom row) datasets.

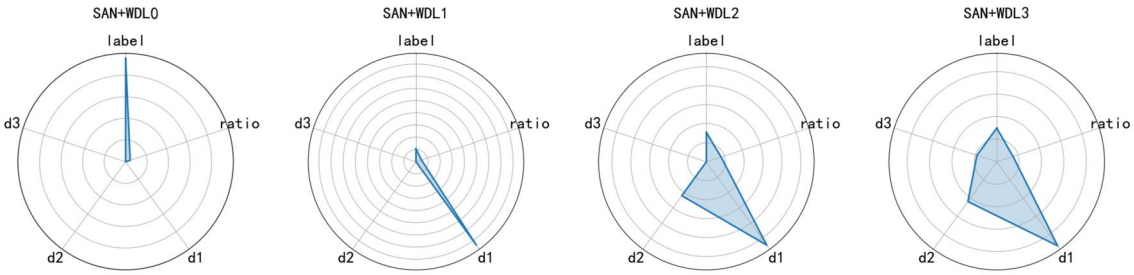

Fig. 5. Weight distribution across difference orders.

Extensibility and future work

The WDL framework is designed as a model-agnostic loss function. Although this paper implements it on BERT, its principles can be extended to other architectures. For instance, in a GNN-based model, WDL could be applied to the final node-level predictions to further refine label relationships beyond what is captured by the graph structure. However, extending WDL to new domains requires careful consideration.

In Extreme Multi-Label Classification (XMLC), where the number of labels can be in the thousands, the direct application of WDL with prompt-based inputs becomes computationally infeasible. A potential solution is a two-stage approach: first, use a candidate-sampling model to retrieve a smaller, relevant subset of labels, then apply WDL to this subset for fine-grained ranking and classification. This would leverage WDL's strength in modeling local dependencies without incurring prohibitive costs.

In Hierarchical Multi-Label Classification (HMLC), the difference calculation could be adapted to respect the hierarchy. For example, differences could be computed primarily among sibling nodes at each level, and perhaps between parent-child nodes, rather than across a flat list. This would allow WDL to model dependencies that are consistent with the predefined label structure. These adaptations, while promising, require substantial future work to validate and implement effectively.

Conclusion

This research introduces the WDL, a novel framework that fundamentally reframes the multi-label classification task from predicting independent probabilities to modeling a dynamic label distribution. By supervising not only label presence but also their relative proportions and rates of change, WDL effectively captures inter-label dependencies without requiring complex architectural modifications or external knowledge graphs.

Our extensive experiments across four diverse datasets demonstrate three key advantages of the WDL framework:

1. Dynamic Relationship Modeling: WDL successfully captures the nuanced, dynamic trends between emotion labels, leading to more robust and accurate predictions, especially in complex scenarios.

2. Implicit Minority Class Boosting: The focus on relative proportions naturally re-allocates model attention to under-represented classes, yielding substantial improvements in minority class F1-scores and overall balanced accuracy.
3. Architecture-Agnostic Simplicity: As a loss-driven innovation, WDL is a lightweight, plug-and-play module that can be easily integrated with various pre-trained models to enhance their performance with minimal overhead.

Our analysis also shows that while the WDL framework can leverage larger backbone models like BERT-Large for potential peak performance gains, this does not guarantee consistent improvement, highlighting that the core benefits stem from the loss design itself. Despite these strengths, our work also highlights areas for future research, including the development of instance-level adaptive weighting to manage performance trade-offs on high-frequency classes and exploring regularization techniques for stable high-order difference learning. The promising results presented here establish WDL as a potent and flexible tool for a wide range of multi-label classification tasks, paving the way for future explorations into more sophisticated dynamic label modeling.

Data availability

The datasets analyzed during the current study are publicly available and can be accessed as described below: ● GoEmotions Dataset: Available at: https://huggingface.co/datasets/google-research-datasets/go_emotions/tree/main/simplified● NLPCC 2018 Task 1 Dataset: Available at: http://tcci.ccf.org.cn/conference/2018/taskdata.php● SemEval Task E-c Dataset: Available at: https://competitions.codalab.org/competitions/17751● Ren-CECPS Dataset: Access to the Ren-CECPS dataset requires contacting the author, Dr. Kang-Xin, at kang-xin@is.tokushima-u.ac.jp.

Received: 20 February 2025; Accepted: 30 June 2025
Published online: 11 July 2025

References

1. Saleema, J.S., Sairam, B., Naveen, S.D., Yuvaraj, K. & Patnaik, L.M. Prominent label identification and multi-label classification for cancer prognosis prediction. In: *TENCON 2012 IEEE Region 10 Conference*, pp. 1–6 (2012). https://doi.org/10.1109/TENCON.2012.6412321
2. Huang, S. et al. Application of label correlation in multi-label classification: A survey. *Appl. Sci.* **14**(19), 9034 (2024).
3. Tsai, C.-P. & Lee, H.-Y. Order-free learning alleviating exposure bias in multi-label classification. *Proc. AAAI Conf. Artif. Intell.* **34**(04), 6038–6045. https://doi.org/10.1609/aaai.v34i04.6066 (2020).
4. Huang, Y., Giledereli, B., Köksal, A., Özgür, A. & Ozkirimli, E. Balancing methods for multi-label text classification with long-tailed class distribution. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics https://doi.org/10.18653/v1/2021.emnlp-main.643 (2021).
5. Pal, A., Selvakumar, M. & Sankarasubbu, M. Magnet: Multi-label text classification using attention-based graph neural network. In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, https://doi.org/10.5220/0008940304940505 (2020).
6. Lango, M. Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study. *Foundations of Computing and Decision Sciences* **44**(2), 151–178. https://doi.org/10.2478/fcds-2019-0009 (2019).
7. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Durand, T., Mehrasa, N. & Mori, G. Learning a deep convnet for multi-label classification with partial labels. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 647-657 https://doi.org/10.1109/cvpr.2019.00074 (2019).
9. Gao, W., Li, S., Lee, S.Y.M., Zhou, G. & Huang, C.-R. Joint learning on sentiment and emotion classification. In: *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*. CIKM13, pp. 1505–1508. ACM, (https://doi.org/10.1145/2505515.2507830 2013).
10. Chong, J. J. Q. & Aryadoust, V. Investigating the effect of multimodality and sentiments on speaking assessments: a facial emotional analysis. *Educ. Inf. Technol.* **28**(6), 7413–7436. https://doi.org/10.1007/s10639-022-11478-7 (2022).
11. Alsheikh, S.S., Shaalan, K. & Meziane, F. Exploring the effects of consumers' trust: A predictive model for satisfying buyers' expectations based on sellers' behavior in the marketplace. *IEEE Access* **7**, 73357–73372 (2019) https://doi.org/10.1109/access.2019.2917999
12. Maharani, W. & Effendy, V. Big five personality prediction based in indonesian tweets using machine learning methods. *International Journal of Electrical and Computer Engineering (IJECE)* **12**(2), 1973 https://doi.org/10.11591/ijece.v12i2.pp1973-1981 (2022).
13. Liao, L. & Wu, Z. Untangling the relationship between work pressure and emotions in social media: auantitative empirical study of construction industry. *Engineering, Construction and Architectural Management* **31**(2), 767–788. https://doi.org/10.1108/ecam-01-2022-0062 (2022).
14. Al Shamsi, A.A. & Abdallah, S. A systematic review for sentiment analysis of arabic dialect texts researches. In: *Proceedings of International Conference on Emerging Technologies and Intelligent Systems: ICETIS 2021* Volume 2, pp. 291–309 (2022). Springer
15. Hsieh, Y.-H. & Zeng, X.-P. Sentiment analysis: An ernie-bilstm approach to bullet screen comments. *Sensors* **22**(14), 5223. https://doi.org/10.3390/s22145223 (2022).
16. Wang, S., Wang, J., Wang, Z. & Ji, Q. Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *IEEE Trans. Multimedia* **17**(12), 2185–2197. https://doi.org/10.1109/tmm.2015.2484966 (2015).
17. Park, H.-M. & Kim, J.-H. Stepwise multi-task learning model for holder extraction in aspect-based sentiment analysis. *Appl. Sci.* **12**(13), 6777. https://doi.org/10.3390/app12136777 (2022).
18. Shu, H., Peng, W., Li, J. & Lee, D. Sentiment and topic analysis on social media. *Proceedings of the 5th Annual ACM Web Science Conference*, 172–181 (2013) https://doi.org/10.1145/2464464.2464512
19. Zhang, Y. & Xie, Y. Emotion analysis system based on skep model. *Business Intelligence and Information Technology*, 632–642 https://doi.org/10.1007/978-3-030-92632-8_59 (2021).
20. Wang, C., Yang, X. & Ding, L. Deep learning sentiment classification based on weak tagging information. *IEEE Access* **9**, 66509–66518 https://doi.org/10.1109/access.2021.3077059 (2021).
21. He, H. & Xia, R. Joint binary neural network for multi-label learning with applications to emotion classification. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) Natural Language Processing and Chinese Computing, pp. 250–259. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99495-6_21 .

22. Bai, W., Wang, J. & Zhang, X. YNU-HPCC at SemEval-2022 task 4: Finetuning pretrained language models for patronizing and condescending language detection. In: Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., Ratan, S. (eds.) Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 454–458. Association for Computational Linguistics, Seattle, United States (2022). https://doi.org/10.18653/v1/2022.semeval-1.61 .
23. Chen, M., Wang, G., Xue, J.-H., Ding, Z. & Sun, L. Enhance via decoupling: Improving multi-label classifiers with variational feature augmentation. In: *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1329–1333 (2021). Institute of Electrical and Electronics Engineers (IEEE)
24. Tang, T., Tang, X. & Yuan, T. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access* **8**, 193248–193256 (2020).
25. Lin, T. Focal loss for dense object detection. arXiv preprint arXiv:1708.02002 (2017)
26. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M. & Zelnik-Manor, L. Asymmetric loss for multi-label classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91 (2021)
27. Geng, X. Label distribution learning. *IEEE Trans. Knowl. Data. Eng.* **28**(7), 1734–1748 (2016).
28. Ghosh, S., R. Menon, R. & Srivastava, S. Lasque: Improved zero-shot classification from explanations through quantifier modeling and curriculum learning. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 7403–7419. *Association for Computational Linguistics*, 7403–7419 (2023). https://doi.org/10.18653/v1/2023.findings-acl.467 .
29. Lafferty, J., McCallum, A. & Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Icml*, vol. 1, p. 3 (2001). Williamstown, MA
30. Li, Y. & Yang, Y. Label embedding for multi-label classification via dependence maximization. *Neural Processing Letters* **52**(2), 1651–1674 (2020).
31. Zhang, M., Cui, Z., Neumann, M. & Chen, Y. An end-to-end deep learning architecture for graph classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018)
32. Ma, Q., Yuan, C., Zhou, W. & Hu, S. Label-specific dual graph neural network for multi-label text classification. 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)*. Association for Computational Linguistics, (2021). https://doi.org/10.18653/v1/2021.acl-long.298 .
33. Inoue, S., Komachi, M., Ogiso, T., Takamura, H. & Mochihashi, D. Infinite scan: An infinite model of diachronic semantic change. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1605–1616. Association for Computational Linguistics*, (2022). https://doi.org/10.18653/v1/2022.emnlp-main.104 .
34. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y. & Han, W. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. arXiv https://doi.org/10.48550/ARXIV.2302.10205 (2023).
35. He, Z.-F., Zhang, C.-H., Liu, B. & Li, B. Label recovery and label correlation co-learning for multi-view multi-label classification with incomplete labels. *Applied Intelligence* **53**(8), 9444–9462. https://doi.org/10.1007/s10489-022-03945-y (2022).
36. Wang, Y., Deng, J., Wang, T., Zheng, B., Hu, S., Liu, X. & Meng, H. Exploiting prompt learning with pre-trained language models for alzheimer's disease detection. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE pp. 1–5 (2023).
37. Zhou, Y., Kang, X. & Ren, F. Prompt consistency for multi-label textual emotion detection. *IEEE Trans. Affect. Comput.* **15**(1), 121–129 (2023).
38. Ruan, Y. & Li, T. Capsule network with label dependency modeling for multi-label emotion classification. *Frontiers in Artificial Intelligence and Applications* https://doi.org/10.3233/faia230494 (2023).
39. Deng, J. & Ren, F. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Trans. Affect. Comput.* **14**(1), 475–486 (2020).
40. Fei, H., Zhang, Y., Ren, Y. & Ji, D. Latent emotion memory for multi-label emotion classification. *Proc. AAAI Conf. Artif. Intell.* **34**, 7692–7699 (2020).
41. Zhu, Y. & Wu, O. Elementary discourse units with sparse attention for multi-label emotion classification. *Knowledge-Based Systems* **240**, 108114 (2022).
42. Ahanin, Z., Ismail, M. A., Singh, N. S. S. & AL-Ashmori, A. Hybrid feature extraction for multi-label emotion classification in english text messages. *Sustainability* **15**(16), 12539 (2023).
43. Huang, C., Trabelsi, A., Qin, X., Farruque, N., Mou, L. & Zaiane, O.R. Seq2emo: A sequence to multi-label emotion classification model. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4717–4724 (2021).
44. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. & Ravi, S. Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)

Author contributions

Q.H. and M.A.A.M. conceptualized the research and conducted the experiments. A.B.A. performed the data analysis and N.A.N. implemented the algorithms. All authors contributed to the writing and editing of the manuscript.

Funding

This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 1020240051) and the Geran Putra (Grant No. GP-IPS/2023/9773800).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.