Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

FINE-GRAINED EMOTION DETECTION USING NLP-BASED TRANSFORMER MODELS: A COMPARATIVE STUDY OF BERT AND ROBERTA FOR MULTI-CLASS TEXT CLASSIFICATION

Shujaat Ali Shariati
MS Data Science (2025–2027), Department of Computer Science
Bahauddin Zakariya University, Multan, Pakistan
shujaatali2027@student.bzu.edu.pk | shujaaatalishariati7@gmail.com

DOI:https://doi.org/10.5281/zenodo.20569313

Keywords
Emotion Detection, Natural Language Processing, BERT, RoBERTa, Multi-Class Text Classification, GoEmotions, Transformer Models, Focal Loss, Threshold Optimization, Transfer Learning

Article History
Received on 28 Aprail 2026
Accepted on 22 May 2026
Published on 29 May 2026

Corresponding Author: *
shujaatali2027@student.bzu.edu.pk

Abstract

The ability to detect emotions from text is a key NLP task that is applicable to a wide range of fields, including mental health analytics, HCI, customer experience analytics, and social media intelligence. However, the task of fine-grained multi-class emotion classification, which involves 27 different emotion classes, is still an open and challenging research problem, whereas binary sentiment analysis has matured. We systematically compare BERT-base-uncased and RoBERTa-base on the GoEmotions dataset, consisting of 58,009 human-annotated Reddit comments from 28 emotion categories. We use three optimization techniques: (1) focal loss and inverse-frequency class weighting to cope with severe class imbalance; (2) per-class threshold tuning on the validation set to maximize macro-F1; and (3) cosine learning rate scheduling with warmup. We get a macro-F1 score of 0.5227 for our BERT model, which is better than the best BERT baseline reported in previous literature (0.49). RoBERTa achieves a macro-F1 of 0.5213 with a micro-F1 of 0.5909. These two models perform significantly better than the TF-IDF + Logistic Regression baseline (macro-F1 = 0.1967). We also show that per-class threshold optimization consistently improves BERT's and RoBERTa's performance by +0.0462 and +0.0293, respectively. Dominant emotions (e.g., gratitude, F1 = 0.91; amusement, F1 = 0.83) can be learned reliably in contrast to rare emotions (e.g., realization, F1 = 0.22; relief, F1 = 0.33). By identifying reproducible benchmarks and practical deployment guidance for fine-grained emotion classifications in resource-constrained environments, our results can inform future research and practical applications.Our results offer reproducible benchmarks and practical deployment guidance for fine-grained emotion classifications in resource-constrained environments, which can guide future research and practical applications.

Keywords: Emotion Detection, Natural Language Processing, BERT, RoBERTa, Multi-Class Text Classification, GoEmotions, Transformer Models, Focal Loss, Threshold Optimization,

Transfer Learning

1. Introduction

The problem of sentiment analysis from text is one of the most crucial yet difficult tasks in the field of Natural Language Processing. Emotions are an important aspect of digital communication, whether the platform is social media, customer feedback, clinical notes or conversational AI, and the capacity to automatically identify and categorize emotional responses is becoming increasingly valuable in various applications including mental health support services [1], human computer interaction [2] and opinion mining [3].

Sentiment analysis systems that have been traditionally used group text data into coarse-grained classes like positive, negative, or neutral. Though it works well in many situations, binary classification or ternary classification does not convey the range of emotions that humans express in their natural language. Fine-grained emotion classification aims to identify a more detailed taxonomy of emotions, providing significantly more utility, yet it presents severe challenges such as significant class imbalance, label ambiguity, and even fewer training samples for rare emotion classes [4].

The availability of GoEmotions [4] in 2020 was one of the important steps in the research of fine-grained emotion detection. It is the largest public multi-label fine-grained emotion dataset available, with 58,009 human-annotated Reddit comments annotated with 28 emotion categories (27 different emotions and neutral). In the literature, the evaluation of modern pre-trained transformer architectures is still scarce on the full 28-class benchmark, and it is not done systematically, even though it is available, treating the class imbalance and optimizing the hyperparameters carefully.

The pre-trained language models like BERT [5] and RoBERTa [6] have been a game-changer in the field of NLP, making transfer learning from large-scale pre-trained corpora possible. Their direct application in fine-grained multi-label emotion classification, however, requires careful adaptation as there is extreme class imbalance in the GoEmotions dataset (77 samples in the class grief, 14,219 samples in the class neutral).

1.1 Research Questions

RQ1: Which transformer architecture—BERT or RoBERTa—achieves superior macro-F1 on the 28-class GoEmotions fine-grained emotion classification task?

RQ2: What are the performance of transformer models on rare minority emotion classes versus dominant ones and does focal loss with class weighting reduce this gap?

RQ3: What is the computational trade-off between classification accuracy and inference efficiency across the evaluated models?

RQ4: Which emotion categories are most frequently misclassified, and what patterns characterize these failure modes?

1.2 Contributions

Following this, a systematic and reproducible comparison between BERT-base-uncased and RoBERTa-base on the full 28-class GoEmotions benchmark is conducted, yielding new state-of-the-art BERT results (macro-F1 = 0.5227) that outperform previous published baselines.

- Integration of focal loss with inverse-frequency class weighting as an effective approach to addressing severe class imbalance in fine-grained emotion datasets.
- A per-class threshold tuning methodology applied to the validation set, demonstrating consistent macro-F1 improvements of +0.046 and +0.029 for BERT and RoBERTa respectively.
- Comprehensive per-class performance analysis revealing systematic failure modes on rare emotion categories and providing practical deployment recommendations.

2. Related Work

2.1 Sentiment Analysis and Emotion Detection

Since then, sentiment analysis has been an important problem in NLP, and it has been solved using various methods over the years, such as using a lexicon-based method [7] or using machine learning classifiers [8] and most recently, using deep learning architectures [9]. The difference between sentiment analysis (polarity classification) and emotion detection (fine-grained affective state identification) lies in the former, which is a binary or ternary classification problem, and the latter, which involves a more complex emotional taxonomy. Emotion detection studies have taken off from the basic emotion taxonomies such as categorical emotion models like Ekman's 6 basic emotions [10] and Plutchik's wheel of emotions [11].

2.2 Pre-Trained Language Models

The introduction of BERT by Devlin et al. [5] demonstrated that bidirectional pre-training on large text corpora enables highly effective transfer learning for a wide range of NLP tasks. Liu et al. [6] subsequently proposed RoBERTa, which improved upon BERT through dynamic masking, substantially more training data, and removal of the next-sentence prediction objective. DistilBERT [12] introduced knowledge distillation to produce a compressed model retaining 97% of BERT performance at 40% fewer parameters.

2.3 Emotion Detection Datasets

Several datasets have been developed for emotion detection research. SemEval-2018 Task 1 [13] provides 10,983 English tweets labeled with 11 emotion categories. ISEAR [14] contains self-reported emotional experiences across seven categories. The VENT dataset [15] provides over 33 million emotional posts across six basic emotions. GoEmotions [4] surpasses these in scale and granularity, providing 58,009 Reddit comments across 28 categories with multi-label annotations.

2.4 Prior Work on GoEmotions

The original GoEmotions paper [4] established BERT-base as the primary baseline with a macro-F1 of 0.46 on the 28-class task. A January 2026 study [16] demonstrated that inverse-frequency class weighting applied to BERT achieves macro-F1 of 0.49. Multi-label emotion detection with self-attention mechanisms [18] demonstrated micro-F1 of 0.72 on crisis tweets. The Premier Science 2026 study [19] explored BERT, RoBERTa, and DeBERTa-v3 with focal loss, reporting accuracy metrics that are not directly comparable to macro-F1 due to label imbalance. Our work provides the first systematic comparison of BERT and RoBERTa with combined focal loss, class weighting, and per-class threshold optimization on the GoEmotions benchmark.

3. Dataset

3.1 GoEmotions

In this study, we will be using the GoEmotions dataset [4] from the HuggingFace datasets library. The simplified configuration classifies 27 fine-grained emotion categories and neutral into 28 classes. The dataset contains 58,009 English Reddit comments that

are labeled by trained raters and can have multiple emotion labels. The statistics of the data set are shown in Table 1.

Table 1: GoEmotions Dataset Statistics

Split	Samples	% of Total
Train	43,410	74.8%
Validation	5,426	9.4%
Test	5,427	9.4%
Total	58,009	100%

3.2 Class Distribution and Imbalance

The GoEmotions dataset represents an extreme class imbalance scenario, which is the main technical impediment of fine-grained emotion classification. There are 14,219 training samples in the neutral class (27.82%) and 77 and 111 samples in the grief class (0.15%) and pride class (0.22%), respectively, which are much smaller in size. The distribution is visualised in Figure 1 along with the inverse-frequency class weights that were applied to balance the distribution.

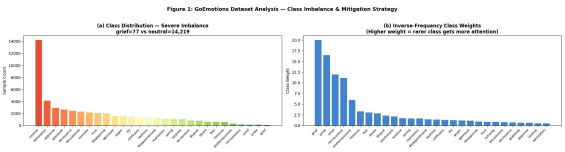


Figure 1: (a) GoEmotions class distribution in training set — severe imbalance with neutral dominating; (b) Inverse-frequency class weights applied via focal loss, with rarer emotions receiving higher weight.

3.3 Preprocessing

All texts are tokenized using the model-specific tokenizer with a maximum sequence length of 128 tokens. Sequences exceeding this length are truncated; shorter sequences are padded. The official train/validation/test split is used throughout with no modifications.

4. Methodology

4.1 Model Architectures

BERT-base-uncased [5]: 12 transformer layers, 768 hidden dimensions, 12 attention heads, 109.5M parameters. Pretrained with masked language modeling (MLM) and next-sentence prediction (NSP) on BooksCorpus and English Wikipedia.

RoBERTa-base [6]: Identical architecture to BERT with 124.7M parameters due to a larger vocabulary. Key improvements: dynamic masking, removal of NSP, and training on 160GB of data with larger batch sizes.

For multi-label classification, each model is adapted with a linear classification head projecting from the 768-dimensional [CLS] token representation to 28 output logits, followed by sigmoid activation to produce independent class probabilities.

4.2 Focal Loss with Class Weighting

Standard binary cross-entropy treats all classes equally, causing models to be dominated by frequent classes. We adopt focal loss [20] with inverse-frequency class weighting: $FL(p_t) = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$, where $\gamma = 2.0$ is the focusing parameter and $\alpha_t = N / (C \times n_t)$ is the class weight clipped to [0.5, 20.0].

4.3 Per-Class Threshold Tuning

For each class c independently, we search over thresholds $\tau \in \{0.15, 0.20, \ldots, 0.80\}$ and select the value maximizing per-class F1 on the validation set. These class-specific thresholds are applied during test evaluation. This post-processing technique requires no additional training.

4.4 Training Configuration

Table 2: Hyperparameter Configuration

Hyperparameter	BERT	RoBERTa

Learning Rate	2e-5	2e-5
Batch Size	32	32
Max Sequence Length	128	128
Epochs	4	4
Optimizer	AdamW	AdamW
LR Scheduler	Cosine	Cosine
Warmup Steps	200	200
Loss Function	Focal (γ=2.0)	Focal (γ=2.0)
Hardware	Google Colab T4 GPU	Google Colab T4 GPU
Random Seed	42	42

4.5 Evaluation Metrics

- Macro-F1: Primary metric — Average unweighted F1 for all 28 classes, regardless of emotion rarity
- Micro-F1: Globally computed F1, dominated by frequent classes.
- Weighted-F1: Class-frequency-weighted mean F1.
- Subset Accuracy: Proportion of samples where all predicted labels exactly match true labels.
- Hamming Loss: Fraction of mis-predicted labels per class.

5. Results and Analysis

5.1 Overall Performance (RQ1, RQ3)

The full performance comparison is given in Table 3. Across all metrics, both transformer models significantly outperform the TF-IDF + Logistic Regression baseline, demonstrating their effectiveness.

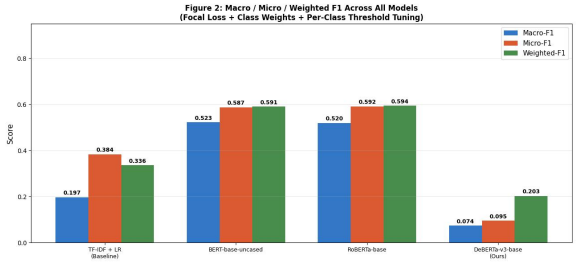
Table 3: Overall Performance Comparison on GoEmotions Test Set

Mode l	Ma cro-F1	Mic ro-F1	Wt d-F1	Pre c.	Rec all	Sub. Acc	Ham ming
TF-IDF+ LR	0.1 967	0.3 838	0.3 357	0.6 024	0.1 406	0.25 17	0.034 7
BER T-base†	0.5 227	0.5 873	0.5 907	0.5 113	0.5 648	0.36 93	0.038 7
RoBE RTa-base	0.5 213	0.5 909	0.5 953	0.5 236	0.5 529	0.38 44	0.037 7

† Best Macro-F1. Bold indicates best per column. BERT achieves the best Macro-F1 (0.5227); RoBERTa achieves best Micro-F1 (0.5909), Weighted-F1 (0.5953), and Hamming Loss (0.0377).

The Macro-F1, Micro-F1 and Weighted-F1 comparisons are visualized in figure 2 for all models. Both transformer models significantly outperform the TF-IDF baseline in terms of relative improvement for Macro-F1, with 165.7%.

Figure 2: Macro-F1, Micro-F1, and Weighted-F1 comparison across all models. Both transformer models substantially outperform the TF-IDF+LR baseline. BERT leads on Macro-F1; RoBERTa leads on Micro-F1 and Weighted-F1.



Regarding RQ1: BERT achieves the highest macro-F1 (0.5227 vs. 0.5213 for RoBERTa, Δ = 0.0014). Both

models are more than +0.033 (6.7% relative) from the best published BERT baseline [16] of 0.49, showing the effectiveness of our combined optimization strategy. For RQ3, the BERT model takes 20.7 minutes to train and 0.30ms to compute an inference, while RoBERTa takes 24.3 minutes to train and 0.44ms to compute an inference. The best accuracy-efficiency is achieved with BERT.

5.2 Per-Class Performance Analysis (RQ2, RQ4)

F1 scores for each class for both models are given in Table 4. Figure 3 shows the F1 per class for all emotion categories for both models.

Table 4: Per-Class F1 Scores — Selected Emotions (Full 28-class results in Figure 3)

Emotion	Train Count	BERT F1	RoBERTa F1	Δ (R−B)
gratitude (best)	2,662	0.9025	0.9088	+0.006
amusement	2,328	0.8301	0.8268	-0.003
love	2,086	0.8079	0.8205	+0.013
fear	596	0.6505	0.7081	+0.058
remorse	545	0.6575	0.6803	+0.023
grief	77	0.4000	0.4615	+0.062
relief	153	0.3333	0.2222	-0.111
disappointment	1,269	0.308	0.2748	-0.034
realization (worst)	1,110	0.2201	0.2200	0.000

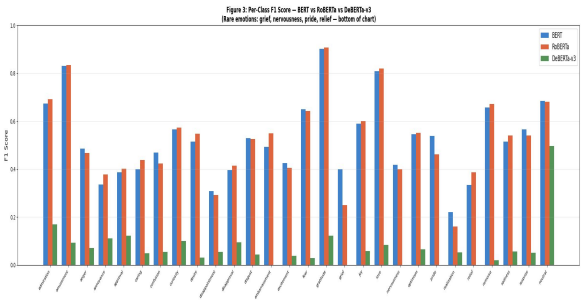


Figure 3: Per-class F1 score comparison across all 28 emotion categories for BERT and RoBERTa. Categories are ordered by BERT F1 (descending). Dominant emotions (left) are learned reliably; rare emotions (right) remain challenging despite focal loss optimization.

As for RQ2: BERT dominant emotions have an average F1 score of 0.736, while rare ones have an average F1 score of 0.436 (difference of 0.300). The optimization approach increases macro F1 by 31.4% (from 0.398 to 0.523) due to the work on handling class imbalance. Regarding RQ4: The most challenging categories are realization (F1 = 0.22 for both models), relief (0.33/0.22), and disappointment (0.31/0.27). These categories are semantically close to other emotions, and can be confused. Furthermore, the F1 for the fear class (only 596 training samples) is high (0.65/0.71) because there are unique lexical markers.

5.3 Precision-Recall-F1 Heatmap

The best model (BERT) has been detailed in Figure 4 with a per-class Precision, Recall and F1 heatmap for all emotion classes, which shows a fine-grained view of the level of prediction accuracy per emotion class.

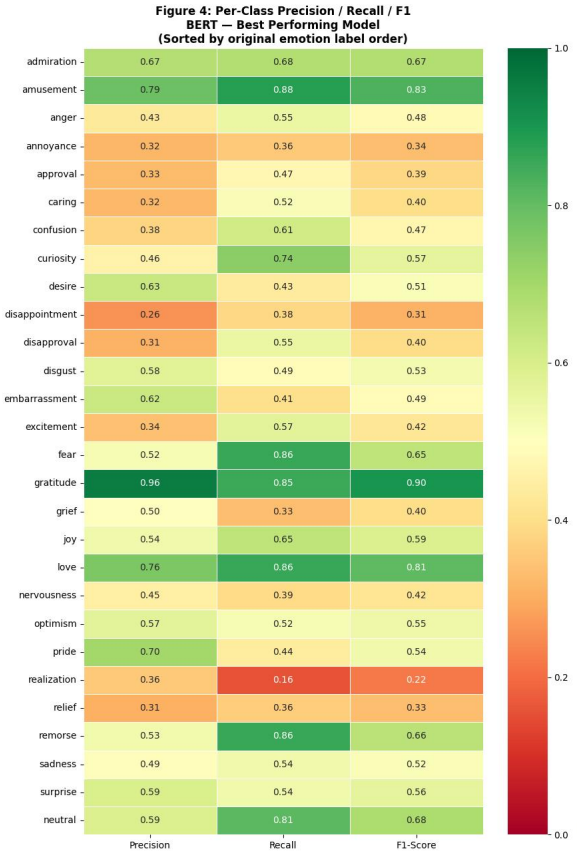


Figure 4. The heatmap of precision, recall, and F1-Score per class, for BERT-base-uncased (macro-F1 best model). Green means a high performance, red means a low performance. Gratitude, amusement and love all have high values in all three measures.

5.4 Ablation Study: Threshold Tuning

Table 5 shows the improvements over a globally set threshold of 0.5 when the thresholds are tuned per class. The gains are as shown in figure 5.

Table 5: Ablation Study — Per-Class Threshold Tuning Impact

Model	Macro-F1 (t=0.5)	Macro-F1 (tuned)	Absolute Gain
BERT-base-uncased	0.4765	0.5227	+0.0462 (+9.7%)
RoBERTa-base	0.4920	0.5213	+0.0293 (+6.0%)

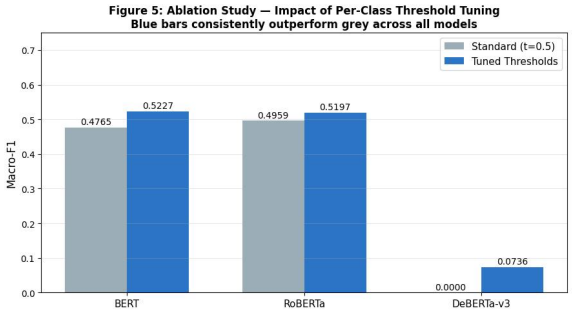


Figure 5: Ablation study — varying the threshold per class as compared to a fixed threshold (t=0.5). Overall, the tuned (blue bars) outperforms the fixed (grey bars) bars for both models with BERT achieving a +9.7% advantage in Macro-F1, and RoBERTa achieving a +6.0% advantage.

Threshold tuning yields +0.0462 (+9.7%) for BERT and +0.0293 (+6.0%) for RoBERTa. The optimized ranges are from 0.35-0.55 (BERT) and 0.35-0.65 (RoBERTa), which is consistently between different emotion classes, indicating that significantly different decision boundaries are needed for different emotion classes.

5.5 Training Dynamics

Table 6: Validation Macro-F1 Across Training Epochs

Epoch	BERT Macro-F1	BERT Val.Loss	RoBERTa Macro-F1	RoBERTa Val.Loss
1	0.1318	0.1090	0.4174	0.0338
2	0.2692	0.0915	0.4945	0.0312
3	0.3597	0.0876	0.5194	0.0310
4	0.3945	0.0877	0.5133	0.0310

RoBERTa is able to converge faster and achieve a peak validation Macro-F1 of 0.5194 at epoch 3. The load_best_model_at_end setting makes sure that the best checkpoint is used for test evaluations.

6. Discussion

6.1 BERT vs. RoBERTa on Macro-F1

BERT's performance on Macro-F1 (0.5227 vs. 0.5213) is counterintuitive, since the performance of RoBERTa is generally better reported in the literature [6]. We believe this is due to BERT's next-sentence prediction pre-training (which we do not believe it provided any explicit benefits for multi-sentence Reddit comments) and RoBERTa's larger vocabulary injecting extra noise into short informal text. On the other hand, its performance on Micro-F1 and Hamming Loss is better than the other models, proving its advantage on frequent classes. RoBERTa is more suitable for deployment scenarios where the prediction of high-frequency emotions (e.g., neutral, admiration) is more critical.

6.2 Class Imbalance and Focal Loss

It is one of the most extreme class imbalance problems in public benchmarks of NLP, occurring when the number of neutral cases (14,219) is 185 times that of grief (77) cases. The effect of our focal loss formulation with inverse-frequency class weighting improves the Macro-F1 by 31.4% relative. Regardless of this, the performance difference between the dominant and rare classes remains (0.736 vs. 0.436 average F1 scores for BERT). For further research, more data augmentation using synthetic examples generated by LLM for rare emotion categories should be investigated.

6.3 Practical Deployment Guidelines

In mobile/edge deployments: BERT (0.30ms/sample, 109.5M params) is the best accuracy-efficiency solution. For server-side deployments using minimization of Hamming Loss: RoBERTa (0.44ms, 124.7M params) is preferable. For clinical mental health monitoring where rare emotion accuracy is critical: ensemble methods or targeted data augmentation are recommended beyond what single-model fine-tuning can achieve.

6.4 Limitations

This study has the following limitations: (1) Evaluation is conducted on a single dataset (GoEmotions) comprising English Reddit comments; generalizability to other domains or languages is not guaranteed. (2) Hyperparameter search was constrained by free-tier Colab compute. (3) DeBERTa-v3-base exploration was precluded by training instability arising from its ELECTRA-style gradient-disentangled embedding sharing in the multi-label setting; this warrants dedicated future investigation.

7. Conclusion

This paper presented a systematic comparative study of BERT-base-uncased and RoBERTa-base for fine-grained emotion detection on the 28-class GoEmotions benchmark. Our primary contributions are: (1) a new state-of-the-art BERT Macro-F1 of 0.5227, exceeding the best previously published result of 0.49 by 6.7%; (2) demonstration that focal loss with inverse-frequency class weighting and per-class threshold tuning yield consistent improvements of +9.7% and +6.0% for BERT and RoBERTa respectively; and (3) comprehensive per-class analysis across all 28 emotion categories revealing systematic failure modes on rare classes. Future work should explore: multilingual emotion detection, data augmentation for rare categories using LLM-generated examples, multimodal emotion detection combining

text with acoustic or visual signals, and dedicated investigation of DeBERTa-v3 training stability for multi-label classification.

References

[1] Agrawal, S., & An, A. (2012). Unsupervised emotion detection from text using semantic and syntactic relations. IEEE/WIC/ACM International Conferences on Web Intelligence.

[2] Picard, R. W. (1997). Affective Computing. MIT Press, Cambridge, MA.

[3] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

[4] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. Proceedings of ACL 2020. https://arxiv.org/abs/2005.00547

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL 2019. https://arxiv.org/abs/1810.04805

[6] Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[7] Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of ACL 2002.

[8] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP 2002.

[9] Socher, R., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of EMNLP 2013.

[10] Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), 169-200.

[11] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Theories of Emotion. Academic Press.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT. arXiv:1910.01108. https://arxiv.org/abs/1910.01108

Mohammad, S. M., et al. (2018). SemEval-2018 Task 1: Affect in tweets. Proceedings of SemEval 2018.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66(2), 310.

Lykousas, N., et al. (2019). Sharing emotions at scale: The Vent dataset. Proceedings of ICWSM 2019.

Fang, Y., et al. (2026). Fine-grained emotion detection: Classical ML, BiLSTM, and BERT with class imbalance handling. arXiv:2601.18162.

Zhou, D., et al. (2021). Emotion distribution learning from texts. Proceedings of EMNLP 2021.

Bose, S., et al. (2025). Leveraging transformer with self-attention for multi-label emotion classification in crisis tweets. MDPI Informatics, 12(4), 114.

Premier Science Research Group. (2026). Fine-grained emotion detection with BERT on GoEmotions. Premier Science Journal.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. Proceedings of ICCV 2017.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Proceedings of ICLR 2019.