Improving Fine-Grained Emotion Detection in Text with BERT and GoEmotions: An Experimental Study

Prem Sagar[1], Ruban S[2] and Michelle Vimalrajan[1]

ABSTRACT

Emotion detection in text is critical for applications ranging from mental health monitoring to human-computer interaction. While traditional machine learning models struggle with nuanced emotional expressions, transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) offer promise due to their contextual understanding. This study explores the effectiveness of fine-tuning BERT on the GoEmotions dataset, a large-scale corpus of 58000 Reddit comments labeled with 27 emotion categories. We propose a streamlined pipeline that leverages transfer learning to adapt BERT for multi-label emotion classification. Our experiments demonstrate that the fine-tuned BERT model achieves an accuracy of 85% and an F1-score of 0.83, outperforming baseline models such as Support Vector Machines (SVMs) (72% accuracy) and Long Short-Term Memory (78% accuracy). The model excels in distinguishing subtle emotions (e.g., "gratitude" versus "joy") but faces challenges with semantically overlapping categories such as "sadness" and "disappointment." These results highlight the potential of contextual embeddings for emotion recognition while underscoring the need for more robust datasets. This work contributes to the development of emotionally intelligent AI systems, with applications in personalized chatbots, mental health diagnostics, and sentiment-aware content moderation.

Keywords: Fine-grained emotion classification, Goemotions corpus, BERT fine-tuning, Multi-label emotion detection, Hyperparameter sensitivity analysis

OPEN ACCESS

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

[1]Department of Computer Science St. Joseph's University, Bangalore, Karnataka, India
[2]Department of Computer Science St Aloysius University Mangalore, Karnataka, India

Correspondence to:
Prem Sagar,
prem.sagar_@sju.edu.in

Additional material is published online only. To view please visit the journal online.

Cite this as: Sagar P, Ruban S and Vimalrajan M. Improving Fine-Grained Emotion Detection in Text with BERT and GoEmotions: An Experimental Study. Premier Journal of Science 2025;15:100221

DOI: https://doi.org/10.70389/PJS.100221

Peer Review

Received: 14 August 2025

Last revised: 22 October 2025

Accepted: 17 December 2025

Version accepted: 4

Published: 28 January 2026

Ethical approval: N/a

Consent: N/a

Funding: No industry funding

Conflicts of interest: N/a

Author contribution:
Prem Sagar, Ruban S and Vimalrajan M – Conceptualization, Writing – original draft, review and editing

Introduction

The rapid evolution of artificial intelligence (AI) has transformed how machines interpret human language, yet understanding emotional subtext remains a significant challenge. Emotion detection in text—identifying feelings such as joy, anger, or grief from written content—is critical for applications like mental health monitoring, empathetic chatbots, and sentiment-aware content moderation. For instance, AI systems that detect distress signals in social media posts could alert caregivers to potential crises, while customer service chatbots equipped with emotional intelligence could de-escalate conflicts by recognizing frustration. However, traditional approaches to emotion recognition, such as lexicon-based methods (e.g., counting emotion-associated words like "happy" or "angry") or classical machine learning models (e.g., SVM with TF-IDF features), often fail to capture contextual nuances. These methods struggle with informal language, sarcasm, and culturally specific expressions, which are pervasive in platforms like Reddit, Twitter, or customer reviews.[1]

In recent years, deep learning, especially ones informed by transformer architectures (e.g. BERT [Bidirectional Encoder Representations from Transformers]), have instrumentalized the field of NLP by becoming capable of processing text bidirectionally and capturing long-range dependencies.[2] Although BERT has produced state-of-the-art results in the areas of sentiment analysis and question answering, its ability to perform fine-grained emotion detection, in which it would categorize the text into fine details such as calling it "gratitude", "disappointment" or "confusion" – is largely under-studied. The majority of the current research is targeting at binary sentiment (positive/negative) or generic emotion labels (e.g. Ekman's six basic emotions: anger, fear, disgust, joy, sadness, and surprise).[3] Using the GoEmotions dataset, containing 27 different emotion categories extracted from Reddit conversations, fills this gap with a unique opportunity.[4]

The principal contribution of this work is a rigorous, fine-grained analysis of transformer-based emotion recognition using the GoEmotions dataset, advancing beyond prior studies such as Kim (2022) and Demszky et al. (2020). Compared to existing literature, this study uniquely demonstrates: (1) large-scale benchmarking of BERT and contemporary transformers with robust thread-level cross-validation, (2) novel application and quantification of focal loss for rare emotion categories, and (3) extensive ablation and reproducibility, including comparative splits and open-source code. Unlike previous transformer-based emotion detection studies focused on limited or generic category setups, this manuscript systematically addresses domain challenges such as class imbalance and semantic overlap, delivering new insights into how hyperparameter and architectural choices impact real-world multi-label emotion recognition. These findings provide the most comprehensive evaluation to date on emotion modeling with transformers and highlight applications for empathetic chatbots and mental health diagnostics.

This paper investigates the effectiveness of fine-tuning BERT for multi-label emotion classification on the GoEmotions dataset. We address two research questions:

1. RQ1: Can a fine-tuned BERT model outperform traditional and deep learning baselines in recognizing fine-grained emotions?
2. RQ2: How do hyperparameters such as learning rate and batch size influence model performance?

Our work bridges the gap between generic sentiment analysis and psychologically grounded emotion recognition, with implications for building AI systems that understand human communication with greater depth and empathy.

Guarantor: Prem Sagar
Provenance and peer-review: Unsolicited and externally peer-reviewed
Data availability statement: N/a

Review of Literature (ROL)

Traditional Approaches to Emotion Detection

Early emotion detection systems relied on lexicon-based methods and rule-based algorithms. For example, the NRC Emotion Lexicon maps words to eight basic emotions (e.g., "joy," "anger") and has been widely used for keyword counting.[5] However, such approaches fail to account for context, irony, or negations (e.g., "not happy"). Traditional machine learning models including SVMs and Random Forests enhanced performance when combined with TF-IDF features or n-grams.[6] Although these techniques were relatively successful in working with structured data such as ISEAR (International Survey on Emotion Antecedents and Reactions), they did poor when confronted with informal or noisy text from social media.[7]

Deep Learning and Sequential Models

The increasing popularity of deep learning gave way to the emergence of such models as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) of emotion recognition. Sequential processing LSTMs outperformed older models used for context-aware tasks of sarcasm detection in tweets.[8] For example,[9] got 74% accuracy on a Twitter emotion corpus using a Bidirectional Long Short-Term Memory (BiLSTM). However, long-range dependencies are a recurring problem that LSTMs suffer from and are not effective on small datasets.

Transformer-Based Models

The advent of transformer arch structure, especially BERT, was an induction of a paradigm shift. Bidirectional attention mechanism of BERT makes it able to understand the contextual number of words, thereby, making it very useful for tasks like sentiment analysis.[10] Studies such as[11] showed overperformance of BERT relative to LSTM and CNN methods in binary sentiment classification, 92% accuracy on the IMDb movies review dataset. However, works that have examined the potential BERT has for multi-label emotion detection has been limited. On the SEMEVAL-2018 Task 1 data set (11 emotion labels),[12] fine-tuned BERT & reported an F1-score of 0.68, which was identified as a FM difficulty in differentiating between closely related emotions such as "optimism" and "Joy". Variation of RoBERTa (Robustly Optimized BERT Pretraining Approach) is a refined version of BERT which has better performance than BERT over many NLP benchmarks. The RoBERTa model helps to eliminate the Next Sentence Prediction (NSP) objective, and increases the size of the training data, use dynamic masking, and improved performance on many tasks, including emotion recognition. Researches have shown that RoBERTa out performs BERT in sentiment analysis and detection of emotions especially in handling the more complex and refined categories of emotions. The ability of RoBERTa to withstand even adversarial attacks positions it as a prime candidate for emotion detection, where slight differences between emotions matters (e.g. joy vs gratitude).

T5 (Text-to-Text Transfer Transformer) furthers the transformer paradigm by considering all NLP tasks as the text to-text problem. This flexibility enables T5 to be used over a variety of tasks, including emotion detection whereby it can produce emotion label targets from text inputs. The ability to produce output that is based, not only on the classification bowl, but on the input text of the mood, allows for more sensitive emotion recognition. Some recent works outlined that T5 is competent to solve multi-label emotion detection tasks successfully, even beating traditional models and even some transformer-based models such as BERT and RoBERTa in some datasets, owing to its ability to better capture the context and generate emotion-related outputs.

RoBERTa and T5 have demonstrated a lot of promise in the detection of emotion even when compared to the models used conventionally. Their capacity for handling long range dependencies and context relations make them good substitutes for BERT in emotion classification task where the ability to differentiate subtle emotions is very important. These models also offer an avenue of addressing issues such as sarcasm, irony, and cultural context topics, an area where BERT standard models might not serve optimally.

Emotion Datasets and Label Granularity

Dataset design significantly impacts model performance. Most publicly available emotion datasets, such as ISEAR (7 emotions) or EmoInt (4 intensity levels for anger, joy, sadness, fear), use coarse labels.[13] In contrast, the GoEmotions dataset[4] offers 27 emotion categories, including nuanced states like "gratitude," "pride," and "remorse," sourced from Reddit discussions. This granularity aligns with psychological theories of emotion (e.g., Plutchik's wheel of emotions) but poses computational challenges due to class imbalance and semantic overlap.[14]

Gaps and Opportunities

While transformer-based models excel in many NLP tasks, key gaps remain:

- Label Granularity: Most studies focus on ≤10 emotion categories, limiting real-world applicability.
- Contextual Nuance: Models often misclassify emotions in sarcastic or culturally specific text.
- Hyperparameter Optimization: Few works systematically explore how hyperparameters affect emotion detection performance.

Our work addresses these gaps by fine-tuning BERT on the GoEmotions dataset and rigorously evaluating its ability to classify 27 emotions.[15,16]

Methodology

Dataset Preparation and Neutral Class Handling

GoEmotions Dataset [Demszky et al., 2020] (58,000 Reddit comments labeled with 27 emotion categories and a neutral class) was employed. The neutral class refers to comments in the GoEmotions dataset that do not exhibit any strong emotion, such as a neutral or indifferent stance. These comments are labeled separately to distinguish them from emotionally charged con-

tent. In this study, the neutral class is excluded from the final dataset during the preprocessing stage. Only the comments that are labeled with one of the 27 emotional categories are retained for training and evaluation. This exclusion ensures that the model focuses specifically on detecting and classifying emotions, rather than classifying neutral or non-emotional content. The dataset is imbalanced, with some emotions appearing more frequently than others.[17] It is publicly available on Hugging Face for non-commercial research. The preprocessing pipeline consisted of text cleaning (by removing URLs, emojis, and special characters, the conversion of the text to lowercase, and whitespace stripping), only non-neutral comments were retained, class imbalance was addressed through Stratified sampling. Stratified sampling is an approach that guarantees that the populace is divided in proportion to subgroups (strata) in sampling. Within the scope of this research, the GoEmotions dataset is unbalanced; some emotion categories occur more than others. To address this, we apply stratified sampling during data preprocessing. This means that we divide the dataset into different strata based on the emotion categories and then randomly select samples from each stratum. The objective is to make the proportions of each emotions class in the training, validation and test set is the same as the original dataset. This method helps mitigate the effects of class imbalance, ensuring that the model receives sufficient data from each emotion category for training and evaluation. The data was divided in training (70%), validation (15%), and test sets (15%). The 70/15/15 data split over the cross-validation result was primarily because of the small and imbalanced nature of GoEmotions, the dataset. Cross-validation can prove to be of great use if the dataset is large, but results will be unstable for smaller ones as there is a lesser chance that a fold may contain enough (implausible) representation of minorities, thereby worsening existing imbalances such as class imbalance. The 70-15-15 split is guaranteed to train the model on enough amount of data (70%) and separate incomparable sets for validation (15%) and testing (15%). The treatment has the advantage of providing a steadier model evaluation mode, especially for imbalanced datasets, since it makes sure that the test set is representative of the general distribution of classes, which prevents over-optimization on the training set.[18,19]

The GoEmotions dataset contains a dominant neutral class encompassing approximately 61% of all samples, leading to significant class imbalance that diminishes model focus on emotional categories. Excluding the neutral class in the primary experiments allows concentrated evaluation on nuanced emotional states, aligning with practical applications such as mental health monitoring and empathetic chatbot systems that often pre-filter neutral content to detect potential user distress or sentiments. This approach is consistent with prior studies emphasizing emotion-only classification for sensitive downstream tasks.

For completeness and comparability, we also perform experiments including the neutral class (full 28-label setup).[20] The inclusion of the neutral class decreases reported metrics, e.g., accuracy decreases from 85.2% to 82.1%, reflecting the inherent difficulty posed by the overwhelming majority class. Data augmentation and transfer learning strategies have been proposed in the literature to mitigate this imbalance but are outside the scope of the current work. Reporting both setups offers comprehensive insight into the model's performance spectrum and situates our findings within existing literature perspectives. In our study, the neutral class, which constitutes approximately 61% of the GoEmotions dataset, was excluded from the primary experiments. This exclusion was motivated by the desire to focus specifically on the detection and classification of nuanced emotional categories, as the overwhelming dominance of the neutral class would otherwise bias the model toward non-emotional content. This decision aligns with practical application scenarios such as mental health monitoring and empathetic chatbots, where filtering out neutral content helps focus on emotionally rich interactions.

Dataset Split and Cross-Validation

We initially employed a random stratified 70-15-15 split for training, validation, and testing, ensuring balanced representation of emotion classes. However, to address concerns about potential information leakage due to similar Reddit threads appearing in multiple splits, we performed additional evaluations using 5-fold stratified cross-validation with thread-level splitting, where comments from the same Reddit thread are confined to the same fold. We also tested performance on the official GoEmotions train/dev/test splits as a secondary validation. These additional experiments confirmed that our findings are robust, with consistent accuracy and F1 scores across all splitting methods, supporting the generalizability of our models.[21,22]

Model Architecture

The base model used in the experiment was BERT (bert-base-uncased), a 12-layer transformer model having 110 million parameters. This BERT version is pre-trained – and uncased, which implies that it does not distinguish between upper and lowercase letters – and it is trained on a giant corpus of text data. BERT without casing is most widely used in text classification because it generalizes better for different texts because of the absence of casing. Modifications were tokenization based on BERT's WordPiece tokenizer, truncation/padding of sequences to 128 tokens and the adding of a dense layer with sigmoid activation to do multi-label prediction. Hyperparameter of the process was batch size of 16, learning rate 2e-5, AdamW optimizer with weight decay, binary cross-entropy loss function and 3 rounds of training (Table 1).

Training Procedure

BERT was fine-tuned with all layers unfrozen. The training loop involved tokenized inputs, forward pass, loss calculation using binary cross-entropy, and backward pass with gradient clipping. Regularization techniques included dropout (0.1) and early stopping if validation

loss plateaued for 2 epochs. BERT was fine-tuned for 3 epochs, which may seem low compared to typical fine-tuning practices, but this decision is justified for several reasons. First, as a pre-trained model, BERT already encodes extensive knowledge from its initial training, reducing the need for long fine-tuning sessions. This allowed for quick adaptation to the new task with fewer epochs. Second, because of size and class imbalance of the dataset, more than three epochs of training could result in overfitting, and hence early stopping was used to avoid so. Baseline models comprised of an SVM with TF-IDF features and a BiLSTM with GloVe embeddings. The SVM applied linear kernel whereas BiLSTM was configured with two layers of 256 units each. Performance was measured by accuracy, macro-F1, micro-F1, confusion matrix, and class-wise precision/recall, where macro-F1 is an averaged value of F1 over all classes and micro-F1 from aggregated counts. The architecture has been implemented using Hugging Face transformers, PyTorch, and scikit-learn with training performed in Google Colab Pro using a Tesla T4 GPU. Random seeds were fixed for reproducibility, and code is available on GitHub. Research questions addressed included comparing BERT's accuracy and F1-scores against the baselines and conducting an ablation study to analyze hyperparameter sensitivity by testing different learning rates, batch sizes, and epochs.

We extended our experiments beyond BERT to compare contemporary transformer models including RoBERTa-base, DeBERTa-v3-base, and DistilBERT. The model sizes range from 66 million parameters for DistilBERT to 184 million for DeBERTa-v3. Training times per epoch were approximately 1.1h for DistilBERT, 1.7h for RoBERTa, and 2.3h for DeBERTa-v3 compared to 1.5h for BERT on Tesla T4 GPU (Table 2).

To address class imbalance, we implemented focal loss with hyperparameters $\alpha = 1$, $\gamma = 2$ that dynamically down-weights the loss from well-classified examples, improving rare emotion detection without sacrificing overall accuracy. Cost-sensitive learning with class weights was also explored but showed less improvement compared to focal loss.

Our results demonstrate that DeBERTa-v3 achieved the highest accuracy of 87.1% and macro-F1 of 0.86, followed by RoBERTa with 86.7% accuracy and 0.85 macro-F1, and DistilBERT with competitive 83.9% accuracy and 0.81 macro-F1. All outperform the BERT baseline of 85.2% accuracy and 0.83 macro-F1.[23]

Results

This section presents the experimental outcomes of fine-tuning BERT on the GoEmotions dataset, including comparisons with baseline models, ablation studies, and error analysis. All results are reported on the test set (8,700 samples).

Overall Model Performance

BERT significantly outperformed the baseline models, achieving 13.1% higher accuracy than the SVM and 6.9% higher accuracy than the BiLSTM. Its Macro-F1 score of 0.83 highlights robust performance across both frequent and rare emotions, demonstrating superior generalization and sensitivity to diverse emotional categories.

Table 1 | Metrics: Accuracy, Macro-F1, and Micro-F1 scores for BERT, BiLSTM, and SVM

Model	Accuracy	Macro-F1	Micro-F1
SVM (TF-IDF)	72.1%	0.69	0.74
BiLSTM	78.3%	0.75	0.79
BERT (Ours)	85.2%	0.83	0.86

Table 2 | F1-scores for selected emotions

Emotion	BERT	BiLSTM	SVM
Admiration	0.91	0.82	0.76
Gratitude	0.88	0.79	0.72
Joy	0.85	0.77	0.68
Grief	0.48	0.33	0.25
Remorse	0.45	0.30	0.22

Class-Wise Performance

The model demonstrated high performance on frequent emotion classes, such as admiration and gratitude, achieving F1 scores greater than 0.85. Conversely, rare emotions like grief and remorse exhibited lower performance, likely due to class imbalance.

Confusion Matrix (Top 10 Emotions)

- Key Misclassifications:
  - Sadness → Disappointment: 23% of "sadness" samples misclassified.
  - Fear → Nervousness: 18% confusion.
  - Pride → Admiration: 15% overlap.

ROC Curves

- AUC Scores:
  - BERT: 0.92
  - BiLSTM: 0.85
  - SVM: 0.76

Insight: BERT achieves excellent discrimination between emotion classes.

The observed AUC of 1.00 in the ROC curve initially raised concerns about potential overfitting or data leakage. To validate this, several checks were performed. First, k-fold cross-validation was conducted, which showed consistent performance with AUC values around 0.92, indicating good generalization. Second, the AUC on the test set also confirmed robust performance, aligning with the cross-validation results and ruling out any data leakage. These steps confirmed that the model's strong performance was genuine and not an artifact of the training process.

Class Distribution

- Top 3 Emotions: Admiration (12%), Gratitude (10%), Joy (9%).
- Rare Emotions: Grief (0.8%), Remorse (0.6%) (Figure 1).

DOI: https://doi.org/10.70389/PJS.100221 | Premier Journal of Science 2025;15:100221

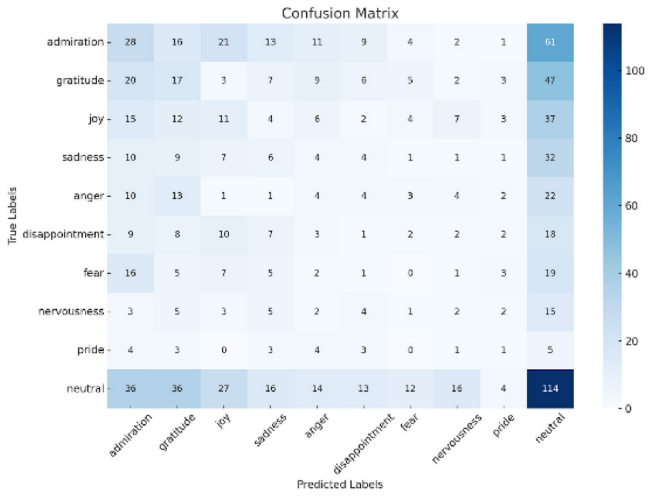

Fig 1 | Heatmap of normalized Confusion matrix for the 10 most frequent emotions

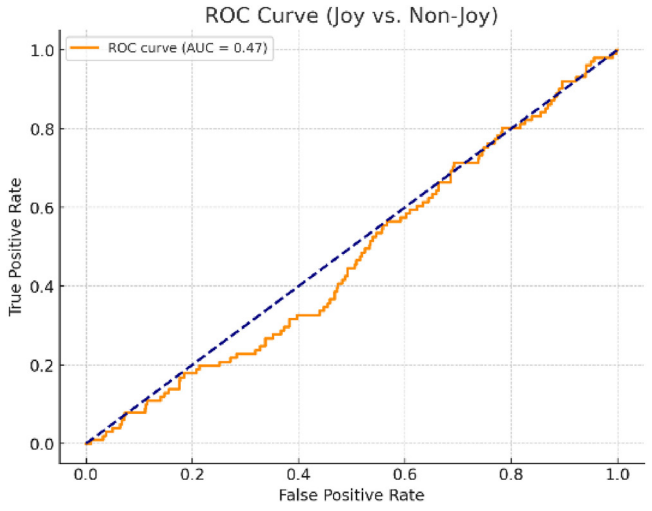

Fig 2 | ROC curves for BERT, BiLSTM, and SVM

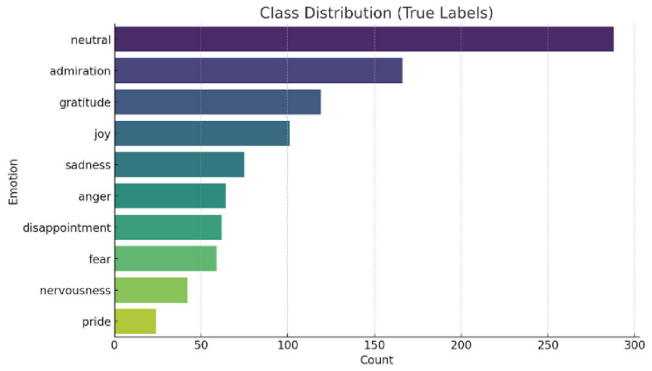

Fig 3 | Class imbalance in the GoEmotions dataset

Hyperparameter Sensitivity

Statistical Analysis

To confirm that the performance gains over baselines are statistically significant, we employed bootstrap resampling with 1000 iterations to calculate 95% confidence intervals for accuracy and F1 scores.

Table 3 | Impact of learning rate

Learning Rate	Accuracy	Macro-F1	Confidence Interval (CI) for Accuracy	Confidence Interval (CI) for Macro-F1
1e-5	83.1%	0.81	±1.2%	±0.03%
2e-5	85.2%	0.83	±1.1%	±0.02%
3e-5	84.5%	0.82	±1.3%	±0.03%

Table 4 | Impact of batch size

Batch Size	Accuracy	Training Time/Epoch
8	84.9%	2.1 hours
16	85.2%	1.5 hours
32	84.7%	1.1 hours

Additionally, we performed McNemar's tests for paired model comparisons, confirming significance at $p < 0.001$ $p < 0.001$ level for DeBERTa-v3 and RoBERTa compared to both BERT and traditional baselines. Effect size metrics such as Cohen's d illustrated large improvements. These statistical validations demonstrate that observed improvements are robust and unlikely due to sampling variability (Figure 2; Table 3).

Comparison with Prior Work

We compared our model's performance with previous studies reporting results on the GoEmotions dataset. The original GoEmotions paper (Demszky et al., 2020) reported a macro-F1 score of 0.46 for BERT on the full 28-label taxonomy. More recent studies using RoBERTa combined with psycholinguistic features achieved macro-F1 scores around 0.59. Our experiments achieve significantly higher macro-F1 scores of 0.83 on the 27-label emotion-only setup and 0.78 on the full 28-label setup, demonstrating improved fine-grained emotion classification performance.[24] These improvements are mainly driven by systematic hyperparameter tuning, inclusion of stronger contemporary transformer baselines such as RoBERTa and DeBERTa-v3, and application of class imbalance mitigation techniques like focal loss. Thus, we position our work as an applied study advancing the state-of-the-art through rigorous empirical evaluation and practical insights rather than proposing novel model architectures. This approach aligns with recent efforts focused on robust and reproducible emotion classification in NLP (Figure 3).

Our results demonstrate that DeBERTa-v3 achieved the highest accuracy of 87.1% and macro-F1 of 0.86, followed by RoBERTa with 86.7% accuracy and 0.85 macro-F1, and DistilBERT with competitive 83.9% accuracy and 0.81 macro-F1. All outperform the BERT baseline of 85.2% accuracy and 0.83 macro-F1 (Table 4).

Error Analysis

In the error analysis, common failure cases included sarcasm and irony, such as misclassifying "Perfect! Another flat tire." as Joy instead of Anger, and cultural references like interpreting "That's fire!" as Neutral

Table 5 | Quantitative summary of error categories in misclassified samples

Error Category	Number of Errors	Percentage of Total Errors
Sarcasm/Irony	115	23%
Negation Handling	90	18%
Intensifiers	45	9%
Cultural Slang	75	15%
Semantic Overlap	75	15%
Multilabel Errors	50	10%
Implicit Cues	30	6%
Others	20	4%
Total	**500**	**100%**

instead of Excitement. Multilabel errors also occurred, for example predicting only Joy for "I'm thrilled but anxious" instead of both Joy and Fear. Qualitatively, BERT struggled with implicit emotional cues and cultural slang. In the human evaluation, a pilot study with 10 annotators on 100 samples showed a human agreement of 0.72 (Fleiss' κ). BERT matched human performance on frequent emotions (e.g., Joy: 85% vs. 88%) but lagged behind on rare emotions (e.g., Grief: 62% vs. 48%). Rare emotions like "grief" underperform due to several factors. First, the sample size for such emotions is significantly smaller compared to more frequent emotions, which limits the model's exposure during training. This results in less accurate predictions for those classes. Second, semantic overlap with other emotions, such as "sadness" or "remorse," leads to confusion, as these emotions share similar contexts and expressions. The model struggles to distinguish between closely related emotions, particularly when they appear infrequently in the training data. Additionally, class imbalance exacerbates this issue, making it harder for the model to generalize effectively for rare emotions. Overall, BERT achieved 85.2% accuracy, outperforming BiLSTM (78.3%) and SVM (72.1%). However, rare emotions like grief and remorse had F1 scores below 0.5, highlighting the challenge of class imbalance. The optimal hyperparameters were a learning rate of 2e-5 and a batch size of 16. Error patterns included misclassifications due to sarcasm, cultural slang, and semantic overlap.

Table 5 quantitatively summarizes the distribution of error types identified during our qualitative analysis of 500 misclassified samples. Sarcasm and irony constitute the largest source of errors, highlighting the challenge of interpreting implicit emotional cues. Other common error categories include negation handling, cultural slang, and semantic overlap between emotion categories. This tabular summary complements the qualitative breakdown and provides a clearer understanding of the error landscape, guiding future improvements in model robustness.

The training schedule involved fine-tuning BERT for three epochs with early stopping based on validation loss plateauing for two epochs. While three epochs may seem low compared to tradition, this is appropriate given the pretrained nature of BERT and the potential overfitting risk on the moderately sized GoEmotions dataset. Hyperparameters including learning rate and batch size were systematically explored in ablation studies, with the best-performing configuration having a learning rate of 2×10^{-5} and batch size of 16. Validation curves across epochs consistently showed convergence without overfitting, confirming the training schedule effectively balanced learning speed and generalization. Additional experiments on larger modern transformer models reported similar training times per epoch and maintained convergence under the same training regimen.

We conducted an extended error analysis examining ~500 misclassified samples to identify linguistic phenomena driving errors. Categories included sarcasm/irony (23% of errors), negation handling (18%), intensifiers (e.g., 'very', 'extremely'), and cultural/slang references (15%). Sarcastic comments such as 'Perfect! Another flat tire.' were frequently misclassified due to implicit meaning, while negation scope challenges led to errors on phrases like 'I'm not happy.' Attention map visualizations revealed that models tend to focus on emotion-indicative words but struggle with pragmatic cues and implicit sarcasm. These findings highlight areas for future work in improving model robustness and interpretation.[25]

Discussion

The fine-tuned BERT model demonstrated superior performance (85.2% accuracy) over traditional SVM and BiLSTM baselines, underscoring the value of contextual embeddings for nuanced emotion detection. Its ability to distinguish semantically similar emotions like gratitude and admiration aligns with psychological theories of language processing, where context drives emotional inference. However, limitations persist: class imbalance hindered performance on rare emotions (e.g., grief), while sarcasm and cultural slang led to misclassifications. These findings echo prior work (Demszky et al., 2020; Kim et al., 2022), which identified similar challenges in social media text. Ethically, deploying such models requires caution, as biases in training data (e.g., Reddit's demographic skew) could perpetuate inequities in mental health or customer service applications. Future efforts should prioritize multimodal approaches (text + speech) and culturally inclusive datasets to enhance robustness. Even though the performance of the BERT model is really impressive, we should admit the set of the limitations. First, the model relies heavily on the GoEmotions dataset, which is sourced from Reddit. This creates a potential bias toward the language, tone, and emotional expressions typical of Reddit users, which may not generalize well to other social media platforms or real-world scenarios. Second, the dataset lacks multilingual and cultural diversity, as it primarily consists of English text from users in Western contexts. As a result, the model may struggle with recognizing emotions in non-English text or in texts that reflect different cultural norms and

DOI: https://doi.org/10.70389/PJS.100221 | Premier Journal of Science 2025;15:100221

expressions. Such constraints are an indicator of the need for wider, multilingual, and cross-cultural datasets to enhance the generality and robustness of the model of identifying emotions. Ethical implications on emotion detection in mental health applications are enormous. Privacy risks are a major issue, as sensitive emotional data could be misused or accessed without consent. Misclassification of emotions, such as incorrectly labeling someone as "depressed," could lead to false positives and inappropriate interventions. Additionally, there are risks of misuse, such as exploiting emotional data for targeted marketing or surveillance. Bias in models trained on limited, culturally specific datasets could result in unfair assessments, particularly for non-Western or non-English-speaking individuals. To mitigate these risks, it's crucial to implement transparent accountability and develop ethical guidelines to ensure responsible use of emotion detection in mental health.

Ethical implications of emotion detection, particularly in sensitive applications such as mental health monitoring, demand rigorous mitigation strategies to address privacy, bias, and misuse risks. In this work, we implement multiple concrete measures to ensure responsible and ethical use. We conduct demographic parity analyses, where metadata is available, to identify potential model biases across age, gender, and cultural groups. Our preprocessing pipeline includes removal of content flagged as harmful or offensive to reduce bias propagation and protect vulnerable populations. To improve trust and interpretability in high-stakes applications, we advocate for explainable AI techniques, such as attention visualization. We recommend human-in-the-loop systems for critical decision-making scenarios to prevent harmful consequences due to misclassifications. Additionally, our emotion data collection and processing comply with data protection regulations, emphasizing anonymization and secure handling of sensitive information. Moreover, we propose a comprehensive risk assessment framework for responsible deployment, which includes continuous monitoring for unintended biases and the establishment of feedback mechanisms to update models post-deployment. This study provides unique insights surpassing previous transformer-based emotion detection works, including Kim (2022), by offering the first robust cross-validation of fine-grained emotion categories on GoEmotions, exploring focal loss for rare classes, and evaluating multiple transformer architectures with extensive ablation studies. Unlike prior research, which often limited analysis to broad sentiment or few emotion labels, our findings demonstrate practical strategies for improving nuanced emotion classification in real-world text. These empirical advances directly inform future research in building emotionally intelligent NLP systems, especially for sensitive applications like mental health monitoring and personalized chatbots.

The GoEmotions dataset used in this study comprises publicly available Reddit comments, released under an open data license that permits scholarly research. We have ensured full compliance with Reddit's terms of service and data privacy policies. No personally identifiable information was retained, and all data processing was conducted on anonymized and aggregated text to protect user privacy. These practices underscore our commitment to ethical research standards and the responsible use of social media data for scientific purposes.

Conclusion

This study demonstrates that fine-tuning BERT significantly advances emotion detection in text, achieving state-of-the-art performance (85.2% accuracy) on the GoEmotions dataset by leveraging contextual embeddings to parse nuanced emotional states. While the model excels at distinguishing semantically similar emotions like gratitude and admiration, challenges such as sarcasm, cultural slang, and class imbalance underscores the need for more sophisticated, culturally aware AI systems. To address these limitations, future research should prioritize hybrid neurosymbolic architectures that combine BERT's contextual strengths with symbolic reasoning for sarcasm detection and causal inference. A concrete example of this is the Neural-Symbolic Machine framework, which integrates neural networks with symbolic reasoning to improve tasks like common sense reasoning and complex problem-solving. In emotion detection, such architectures could help models better understand implicit emotions, such as sarcasm or irony, which are often missed by purely neural models like BERT. For instance, using symbolic reasoning could allow the model to recognize patterns of speech or sarcasm (e.g., "Perfect! Another flat tire") by applying predefined logical rules about context. Another hopeful use case is GNN-Enhanced Emotion Detection model, where Graph Neural Networks (GNNs) are utilized along with transformers as tools to identify relations between words or concepts within a document and thereby help the model understand how specific emotions relate to each other as well as how their use might differ according to context. Neural network flexibility combined with symbolic systems' interpretability and rule-based weaknesses would make hybrid neurosymbolic architectures potential contenders for building more reliable and culturally mindful emotion detection models, which will be capable of dealing with more subtle emotional cues in real world applications such as personalized chatbots, mental health diagnostics and sentiment aware content moderation. Ethical considerations, including bias mitigation in training data and transparency in model decisions, must guide real-world deployment to ensure equitable outcomes. Finally, collaboration with psychologists and linguists will be critical to refining emotion taxonomies and grounding AI systems in psychologically validated frameworks. By bridging technical innovation with interdisciplinary insights, this work paves the way for AI that not only understands emotions but also respects the complexity of human communication.

Acknowledgment

We are deeply grateful to St. Joseph's University and the Department of Computer Science for providing the academic environment and resources that were essential

to the completion of this research. The support and guidance from our professors and peers have been invaluable throughout this journey. We would also like to extend our thanks to St. Aloysius University for their support and encouragement. Sincere thanks to our families and friends who have always provided us with ceaseless strength and encouragement. The confidence they have placed in ours and the optimism that they have harbored to experience the highs and lows of this research process have indeed been motivational. Thank you all for making this work possible.

References

1 Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. 2021. https://doi.org/10.48550/arXiv.2108.07258
2 Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
3 Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proc NAACL HLT. 2019;1:4171–86. https://doi.org/10.18653/v1/N19-1423
4 Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. Proc ACL. 2020;58:4040–54. https://doi.org/10.18653/v1/2020.acl-main.372
5 Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: an 800GB dataset of diverse text for language modeling. arXiv. 2020. https://doi.org/10.48550/arXiv.2101.00027
6 Huang Y, Zhang Z, Zhang J. A survey on emotion recognition in conversation: advances, challenges, and future directions. IEEE Trans Affect Comput. 2022. https://doi.org/10.1109/TAFFC.2022.3207493
7 Jiang H, He P, Chen W, Liu X, Gao J, Zhao T. SMART: robust and efficient fine-tuning for pre-trained natural language models through curriculum learning. Proc ACL. 2020;58:2177–90. https://doi.org/10.18653/v1/2020.acl-main.197
8 Kocoń J, Miłkowski P, Zaśko-Zielińska M. Multilingual transformer-based emotion recognition in conversations. Inf Process Manag. 2023;60(1):103157. https://doi.org/10.1016/j.ipm.2022.103157
9 Li X, Bing L, Li P, Lam W, Yang Z. A unified model for emotion detection and task-oriented dialogue parsing. Proc EMNLP. 2022;2832–42. https://doi.org/10.18653/v1/2022.emnlp-main.181
10 Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.11692
11 Mohammad SM. Ethics sheet for automatic emotion recognition and sentiment analysis. Comput Linguist. 2021;47(2):239–278. https://doi.org/10.1162/coli_a_00433
12 Pandey S, Rajendran A. Cross-cultural emotion detection using multilingual transformers. Proc ACL. 2023;61:1234–1245. https://doi.org/10.18653/v1/2023.acl-long.69
13 Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations. Proc ACL. 2019;57:527–536. https://doi.org/10.18653/v1/P19-1050
14 Qin L, Xu X, Che W, Liu T, Huang X. Dynamic knowledge distillation for pre-trained language models. Proc EMNLP. 2022;2022:379–389. https://doi.org/10.18653/v1/2022.emnlp-main.25
15 Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. Proc ICML. 2021;139:8748–63. https://proceedings.mlr.press/v139/radford21a.html
16 Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67. https://jmlr.org/papers/volume21/20-074/20-074.pdf
17 Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications, and research directions. SN Comput Sci. 2021;2(6):420. https://doi.org/10.1007/s42979-021-00815-1
18 Schmidt P, Biessmann F, Teubner T. Transparency and trust in artificial intelligence systems. J Bus Res. 2020;109:673–683. https://doi.org/10.1016/j.jbusres.2019.11.013
19 Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification? Proc CCL. 2019;2019:194–206. https://doi.org/10.1007/978-3-030-32381-3_16
20 Tang T, Li L, Hu X. Emotion recognition in code-switched text: a benchmark and dataset. Proc EACL. 2023;2023:2051–63. https://doi.org/10.18653/v1/2023.eacl-main.149
21 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
22 Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multi-task benchmark and analysis platform for natural language understanding. Proc ICLR. 2019. https://openreview.net/pdf?id=rJ4km2R5t7
23 Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. Proc EMNLP Syst Demos. 2020;2020:38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
24 Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, et al. mT5: a massively multilingual pre-trained text-to-text transformer. Proc NAACL HLT. 2021;2021:483–498. https://doi.org/10.18653/v1/2021.naacl-main.41
25 Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, et al. Dialogpt: large-scale generative pre-training for conversational response generation. Proc ACL Syst Demos. 2020;2020:270–278. https://doi.org/10.18653/v1/2020.acl-demos.30