Improving Depression Emotion Classification Using RoBERTa

Anoushka Yadav, Kabir Gupta, Garvv Chadha, Krish Gupta
Computational Social Science

***Abstract*—Depression is a complex and heterogeneous mental health condition that manifests through a range of overlapping emotional states rather than a single, clearly defined signal. While prior computational approaches have largely framed depression detection as a binary classification problem, such formulations fail to capture the nuanced and co-occurring emotional expressions commonly observed in social media discourse. In this work, we present a methodologically grounded approach to multilabel depression emotion classification using the RoBERTa transformer architecture. Leveraging the DepressionEmo dataset, which consists of long-form Reddit posts annotated with eight clinically relevant depression-related emotions, we introduce a training framework that combines class-balanced loss with per-label threshold optimization. These methodological enhancements are explicitly designed to address severe class imbalance and emotion-specific precision-recall trade-offs inherent to real-world mental health data. Empirical evaluation demonstrates that the proposed approach consistently outperforms strong transformer baselines, achieving an F1-Macro score of 0.80 and an F1-Micro score of 0.84. Further analyses using ROC-AUC, PR-AUC, confusion matrices, and correlation heatmaps indicate that the model not only improves predictive performance but also learns meaningful relationships between emotional categories. The results underscore the importance of fine-grained emotion modeling for computational social science research and highlight the potential of such systems for mental health monitoring and early intervention.**

***Index Terms*—Depression detection, multilabel classification, emotion analysis, RoBERTa, social media, mental health**

I. Introduction

Depression is among the most prevalent mental health disorders globally and represents a major public health concern due to its association with reduced quality of life, impaired functioning, and increased risk of self-harm. With the widespread adoption of social media platforms, individuals increasingly articulate their emotional experiences and psychological distress online, creating unprecedented opportunities for computational analysis of mental health-related signals. Consequently, natural language processing (NLP)-based approaches have emerged as promising tools for large-scale mental health monitoring.

However, much of the existing literature frames depression detection as a binary classification task, labeling individuals or posts as either depressed or non-depressed. While such approaches may be useful for coarse screening, they fundamentally oversimplify the phenomenology of depression. Clinical psychology literature consistently emphasizes that depression is characterized by a constellation of emotional states-such as hopelessness, worthlessness, sadness, loneliness, and suicidal ideation-that frequently co-occur and vary in intensity across individuals and time.

This mismatch between clinical understanding and computational modeling motivates the need for multilabel emotion classification frameworks that can capture the richness of depressive expression. Rather than predicting a single label, such systems aim to identify multiple co-existing emotional states within a single text instance. In this work, we focus on improving multilabel depression emotion classification by leveraging the RoBERTa transformer architecture and introducing targeted methodological enhancements that address key challenges in this domain.

II. Motivation: Emotion-Level Analysis in Depression Detection

Emotion-level analysis offers a substantially more informative and actionable representation of depressive expression than binary classification. Different emotions correspond to different levels of psychological risk and may necessitate distinct forms of intervention. For example, expressions of sadness or loneliness may indicate emotional distress, whereas explicit mentions of suicide intent constitute high-risk signals requiring immediate attention.

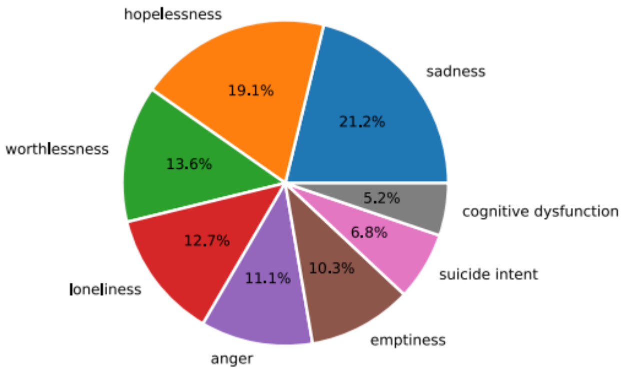
Social media text further complicates this task due to its informal style, frequent use of metaphor, implicit emotional cues, and long narrative structure. Users often describe their experiences indirectly, blending multiple emotional states within a single post. As a result, models that are constrained to a single output label are inherently ill-suited to capture the complexity of real-world depressive discourse.

By adopting a multilabel formulation, emotion-level models can provide a more faithful representation of psychological states, enabling fine-grained analysis and prioritization of critical emotions. Such granularity is particularly valuable in computational social science and mental health applications, where interpretability and downstream utility are as important as predictive accuracy.

III. The DepressionEmo Dataset

A. Dataset Overview

The DepressionEmo dataset is a pioneering resource designed specifically for multilabel classification of depression-related emotions. It consists of 6,037 long-form Reddit posts collected from mental health-related communities. Unlike short tweets, these posts often contain detailed narratives, making them suitable for studying complex emotional expression.



B. Annotation Methodology

Emotion labels were generated using a majority-voting scheme over zero-shot predictions from pretrained language models. These automatically generated labels were subsequently validated by human annotators and ChatGPT, resulting in acceptable interrater reliability. This hybrid annotation strategy balances scalability with quality control, a key consideration in computational social science research.

C. Emotion Categories and Challenges

The dataset includes eight emotion labels: hopelessness, worthlessness, loneliness, sadness, guilt, suicide intent, anxiety/cognitive dysfunction, and emptiness. The multilabel nature of the dataset reflects real-world emotional co-occurrence but also introduces challenges such as severe class imbalance and overlapping linguistic patterns between related emotions.

IV. Related Work and Baseline Models

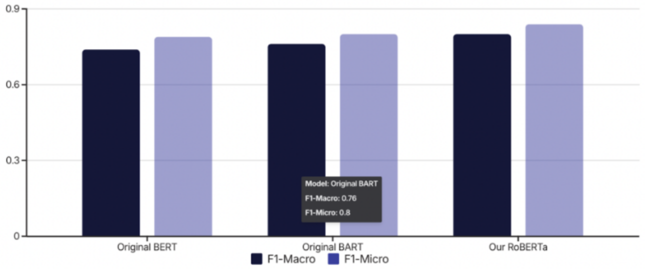
Prior research on depression detection has explored traditional machine learning approaches as well as transformer-based architectures such as BERT and BART. While these models demonstrated the feasibility of emotion-aware classification, their performance was constrained by fixed decision thresholds and limited handling of rare labels.

Among the evaluated baselines, BART achieved the strongest performance with an F1-Macro score of 0.76 and an F1-Micro score of 0.80. However, error analysis revealed that these models struggled particularly with low-frequency emotions such as suicide intent and cognitive dysfunction, motivating the methodological enhancements proposed in this work.

V. Proposed Methodology

A. Model Architecture

The proposed classification framework is built upon RoBERTa-base, a transformer-based encoder architecture that has demonstrated strong performance across a wide range of natural language understanding tasks. RoBERTa-base consists of 12 transformer layers, each with 768-dimensional hidden representations and 12 self-attention heads, enabling the model to capture both local and long-range contextual dependencies within text.



Given an input Reddit post, the text is first tokenized using the RoBERTa tokenizer and passed through the encoder to produce contextualized embeddings for each token. For sequence-level representation, we utilize the final hidden state corresponding to the special [CLS] token, which is commonly adopted as a holistic summary of the entire input sequence in transformer-based classification tasks. This representation encodes aggregated semantic and emotional information from the full post, making it suitable for downstream multilabel prediction.

To mitigate overfitting-particularly important given the moderate dataset size and high model capacity-the [CLS] embedding is passed through a dropout layer with a dropout probability of 0.4. This regularization step randomly deactivates a subset of neurons during training, encouraging the model to learn more robust and generalized feature representations. Finally, a fully connected linear layer maps the regularized embedding to an 8-dimensional output vector, where each logit corresponds to one depression-related emotion category. These logits represent unnormalized confidence scores that are subsequently processed during loss computation and thresholding.

B. Advantages of RoBERTa for Emotion Classification

RoBERTa offers several architectural and pretraining advantages over earlier transformer models such as BERT, which are particularly relevant in the context of emotion classification from social media text.

First, RoBERTa employs dynamic masking, wherein different subsets of tokens are masked during each training epoch. This contrasts with BERT's static masking strategy and allows the model to observe a more diverse set of token prediction contexts during pretraining, resulting in richer contextual representations.

Second, RoBERTa is trained on substantially larger and more diverse corpora, enabling improved language understanding and robustness to linguistic variation. This is especially important for Reddit data, which frequently contains informal language, slang, spelling variations, and emotionally charged expressions.

Third, RoBERTa removes the next-sentence prediction (NSP) objective used in BERT, allowing the model to focus entirely on token-level and sequence-level representation learning. This design choice has been shown to improve performance on downstream classification tasks, particularly those involving longer text sequences.

Collectively, these characteristics make RoBERTa particularly well-suited for modeling the subtle and context-dependent emotional cues present in mental health–related social media discourse.

VI. Training Strategy

A. Multilabel Loss Function

The emotion classification task is formulated as a multilabel learning problem, where each input text can simultaneously belong to multiple emotion categories. Consequently, the prediction of each emotion is treated as an independent binary classification task rather than a mutually exclusive multiclass decision.

To support this formulation, the model is trained using Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss). This loss function combines a sigmoid activation and binary cross-entropy into a single, numerically stable operation. For each emotion label, the sigmoid function converts the corresponding logit into a probability value in the range [0, 1], representing the model's confidence that the emotion is present in the input text.

The overall loss is computed as the sum (or mean) of the binary cross-entropy losses across all eight labels. This approach allows the model to learn label-specific decision boundaries while maintaining independence between emotion predictions, which is appropriate given the overlapping nature of depressive emotional states.

B. Class-Balanced Loss for Imbalanced Labels

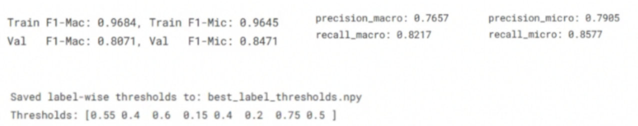
A key challenge in depression emotion classification is severe class imbalance, where certain emotions-such as sadness or hopelessness-occur frequently, while others-such as suicide intent or cognitive dysfunction-are comparatively rare. Without corrective measures, standard loss functions tend to bias the model toward majority classes, leading to poor recall for rare but clinically critical emotions.

To address this issue, we incorporate label-specific positive weighting within the BCEWithLogitsLoss framework. For each emotion label, a positive weight is computed as:

$$\text{Pos\_weight}_i = \frac{N_{\text{negative},i}}{N_{\text{positive},i}}$$

where $N_{\text{positive},i}$ and $N_{\text{negative},i}$ denote the number of positive and negative samples for the i-th emotion, respectively.

This weighting scheme increases the penalty associated with misclassifying positive instances of rare emotions, effectively encouraging the model to pay greater attention to underrepresented labels during training. As a result, the model is less likely to ignore low-frequency but high-impact emotional states, leading to improved recall and more balanced performance across emotion categories.



```
Train F1-Mac: 0.9684, Train F1-Mic: 0.9645     precision_macro: 0.7657     precision_micro: 0.7905
Val   F1-Mac: 0.8071, Val   F1-Mic: 0.8471     recall_macro: 0.8217        recall_micro: 0.8577

Saved label-wise thresholds to: best_label_thresholds.npy
Thresholds: [0.55 0.4  0.6  0.15 0.4  0.2  0.75 0.5 ]
```

VII. Threshold Optimization

In multilabel classification, model outputs are continuous probabilities that must be converted into binary predictions through thresholding. A common but suboptimal practice is to apply a fixed threshold (e.g., 0.5) across all labels. However, this approach fails to account for label-specific characteristics such as prevalence, linguistic ambiguity, and differing costs of false positives versus false negatives.

To overcome this limitation, we employ per-label threshold optimization using a held-out validation set. For each emotion label, decision thresholds ranging from 0.1 to 0.9 are systematically evaluated. At each threshold value, predictions are generated and the corresponding F1-score is computed for that label.

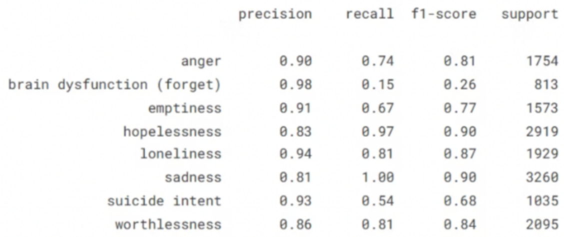
The threshold that maximizes the F1-score is selected as the optimal operating point for that emotion. This process allows each label to adopt a decision boundary that best balances precision and recall according to its statistical and semantic properties. For instance, rare and high-risk emotions such as suicide intent may benefit from lower thresholds that favor recall, whereas more common emotions may require higher thresholds to maintain precision.

This adaptive thresholding strategy plays a critical role in the overall performance of the model and represents a key methodological contribution of this work, demonstrating that post-training optimization can substantially enhance multilabel classification outcomes.

VIII. Experimental Results

The proposed RoBERTa-based model demonstrates consistent and substantial improvements over prior baselines. Overall performance is evaluated using both F1-Macro and F1-Micro scores, which capture complementary aspects of multilabel classification performance. The model achieves an F1-Macro score of 0.80, indicating balanced performance across all emotion labels, and an F1-Micro score of 0.84, reflecting strong overall predictive accuracy.

Label-wise analysis reveals particularly strong performance for emotions such as hopelessness, sadness, loneliness, and worthlessness, with F1-scores ranging approximately from 0.84 to 0.90. These emotions are more frequently represented in the dataset and exhibit clearer linguistic cues, allowing the model to learn robust decision boundaries. In contrast, cognitive dysfunction (forgetfulness) remains a challenging category, with notably low recall. This difficulty can be attributed to both the scarcity of training examples and the subtle, overlapping language patterns associated with this emotion.



```
                           precision    recall  f1-score   support

                   anger       0.90      0.74      0.81      1754
brain dysfunction (forget)     0.98      0.15      0.26       813
              emptiness       0.91      0.67      0.77      1573
           hopelessness       0.83      0.97      0.90      2919
             loneliness       0.94      0.81      0.87      1929
                sadness       0.81      1.00      0.90      3260
         suicide intent       0.93      0.54      0.68      1035
         worthlessness       0.86      0.81      0.84      2095
```

Receiver Operating Characteristic (ROC-AUC) scores exceed 0.87 for all emotion categories, indicating that the model is effective at separating positive and negative instances across labels. Precision-Recall (PR-AUC) analysis provides a more nuanced view under class imbalance, highlighting that while separability remains high for rare emotions, maintaining precision at higher recall levels is inherently difficult. These results validate the importance of class-balanced loss and threshold tuning in achieving stable multilabel performance.

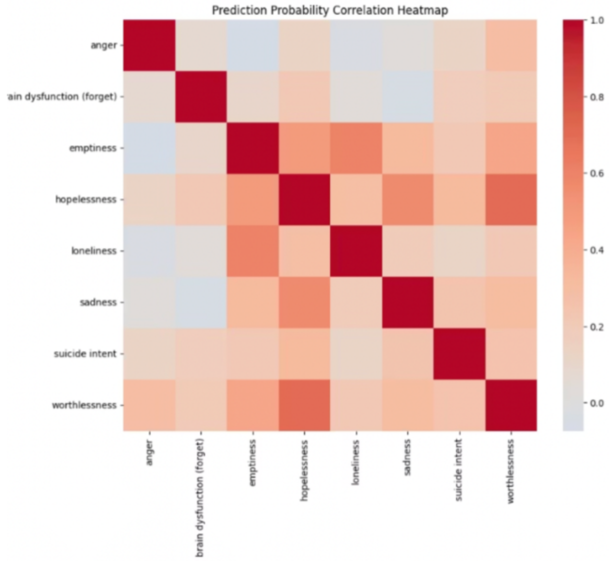
IX. Error Analysis and Correlation Studies

To better understand the strengths and limitations of the proposed model, we conduct a detailed error analysis using confusion matrices, ROC-AUC/PR-AUC trends, and emotion correlation heatmaps. This analysis provides insight not only into predictive performance but also into how the model internalizes relationships between different emotional states.

Confusion matrix analysis indicates that hopelessness and sadness are among the easiest emotions for the model to detect, exhibiting high true positive rates and relatively low false negative counts. These emotions are both well-represented in the dataset and characterized by explicit linguistic cues (e.g., expressions of despair, loss of motivation, or persistent low mood), allowing the model to learn stable decision boundaries. Similarly, emotions such as loneliness and emptiness demonstrate strong performance, though with moderate confusion arising from overlap with sadness-related language.

In contrast, suicide intent presents a more complex error profile. While the model achieves high recall for this critical category, it also produces a non-trivial number of false positives, primarily due to linguistic overlap with closely related emotions such as hopelessness and worthlessness. This behavior reflects a realistic trade-off: language expressing suicidal ideation often shares vocabulary with broader expressions of despair, making strict separation inherently difficult. From an application standpoint, such false positives may be preferable to false negatives, particularly in early warning or triage systems.

Correlation heatmaps further illuminate the structure of emotional co-occurrence. Ground-truth correlations reveal strong positive relationships between hopelessness and worthlessness, as well as between loneliness and sadness, aligning with established psychological theory on depressive symptom clusters. Conversely, anger exhibits weak or near-zero correlation with most other emotions, suggesting that it represents a linguistically and psychologically distinct mode of expression within depressive discourse.



Importantly, the model's predicted correlation patterns closely mirror these ground-truth relationships. This alignment indicates that the classifier is not merely optimizing label-wise accuracy in isolation, but is instead learning meaningful dependencies between emotional categories. Such behavior suggests that the model captures underlying emotional structure rather than relying on spurious lexical cues, strengthening confidence in its interpretability and robustness.

X. Social Impact, Limitations, and Future Work

The findings of this study have significant implications for both computational social science research and applied mental health technologies. By moving beyond binary depression detection toward fine-grained, multilabel emotion classification, the proposed approach enables a more nuanced understanding of psychological states expressed on social media. This granularity is particularly valuable for large-scale mental health monitoring, where different emotional signals may correspond to varying levels of urgency and intervention.

In practical terms, improved detection of high-stakes emotions such as suicide intent and hopelessness can support early identification of individuals at elevated risk, enabling timely human-in-the-loop interventions. At the same time, accurate recognition of lower-intensity but persistent emotions such as loneliness or emptiness can inform longer-term support strategies. From a methodological perspective, this work demonstrates how careful loss design and threshold optimization can substantially enhance transformer-based models in socially sensitive, imbalanced settings.

Despite these contributions, several limitations must be acknowledged. First, although the DepressionEmo dataset is carefully constructed, its overall size remains modest relative to large-scale NLP benchmarks, which may constrain generalization. Second, annotation noise is an inherent challenge in social media-derived mental health data, even when human validation is employed. Third, ethical considerations surrounding privacy, consent, and potential misuse of automated mental health systems remain critical concerns and necessitate cautious deployment. Finally, the current study focuses on a single platform (Reddit), limiting conclusions about cross-platform or cross-cultural generalizability.

Future work may address these limitations through several avenues. Exploring larger and more expressive transformer architectures (e.g., RoBERTa-large or DeBERTa) may further improve performance on subtle and low-frequency emotions. Cross-dataset and cross-platform evaluations would provide stronger evidence of robustness and generalizability. Additionally, incorporating temporal modeling could enable analysis of emotional trajectories over time, offering deeper insight into the progression of depressive states. Finally, user-level longitudinal analysis represents a promising direction for aligning computational models more closely with personalized mental health assessment and intervention.

XI. Author Contributions

All authors contributed substantially and equally to this work. Anoushka Yadav was responsible for the design and implementation of the RoBERTa-based classification model, conducted training experiments, performed threshold optimization, and led quantitative analysis of results. Kabir Gupta contributed to the conceptual design of the study, assisted in defining the multilabel emotion formulation, supported the evaluation strategy and metric selection, and participated in result interpretation and report development. Garv Chadha carried out model evaluation, generated performance visualizations including ROC–AUC, PR–AUC, confusion matrices, and correlation heatmaps, and contributed to error analysis and interpretation. Krish Gupta contributed to methodological design, supervised the experimental setup, assisted in refining model architecture and loss formulation, and provided critical revisions to improve clarity and technical rigor. All authors jointly contributed to problem formulation, discussion of results, writing the report, and final review.

References

[1] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

[2] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[3] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," EMNLP, 2020.

[4] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.

[5] DepressionEmo Dataset: A Novel Dataset for Multilabel Classification of Depression Emotions.