A Hybrid BERT-RoBERTa Ensemble for Robust Classification of Emotion Classes

1st Mohit Jaiswal
B.Tech, CSE
Bennett University
Greater Noida, India
mohitjaiswal2507@gmail.com

Abstract—**Text Emotion Classification (TEC) is a challenging Natural Language Processing task, especially when it comes to with rich nuances, context, and sarcasm. While models like BERT and RoBERTa have made very strong individual performances. they are mostly restricted to training on 5-7 basic emotions, not This captures complex states such as confusion or sarcasm. Paper proposes a hybrid architecture: fine-tuning of `bert-base-uncased` and `roberta-base` simultaneously. We concatenate the feature embeddings from both models to create a richer 1536-dimensional feature vector. The model is trained on a comprehensive 13-label dataset, including but not limited to sarcasm, disgust, confusion, and shame. We show that this hybrid model outperforms all of the single models baselines. which will allow a much stronger solution to subtle, real-world emotion recognition.**

Index Terms—**Text Emotion Classification, BERT, RoBERTa, Ensemble Learning, Sarcasm Detection, Transformer, Deep Learning, Nuanced Emotions.**

I. Introduction

Text Emotion Classification (TEC) is an important task in Natural Language Processing, the task of automatically classifying and categorizing human emotions found in text. Its applications are as far-reaching as analyzing customer feedback and monitoring public sentiment on social media to Building Pervasive empathy in human-computer interaction systems [1], [7]. However, traditional sentiment analysis focuses on simple positive/negative polarity, the TEC is far more complex. Human languages are full of ambiguities, context-dependent meaning and figurative language. One of the main obstacles has been that most models are trained on basic emotions, that is, those of joy, anger, and fear, while failing to represent the more nuanced and subtle human states such as confusion, shame, or sarcasm. This Paper investigates the usage of advanced Transformer models on a This highly granular 13-label dataset overcomes these challenges.

A. Real-world example

Arguably, the most challenging part about TEC is context and sarcasm. Consider the sentence: `"Oh, great. The flight is delayed again."` A naive model which was trained only on the basic emotions would see the positive-valence word "great" and likely classify the emotion as **Joy**. This is demonstrably wrong. The context provided by ”flight is delayed” and ”again” are completely inverted the word's meaning, making it into an expression of **Anger** or **Sadness**. Herein lies the critical difference between literal meaning (lexical sentiment) and intended emotional context. An effective classifier needs to know more than just context; also be trained on a dataset which explicitly includes these more complex, nonliteral emotion categories, such as **Sarcasm**.

II. General Overview

The present paper describes our solution to the challenges of nuanced TEC. This paper is organised according to the required structure, where we begin with an introduction to the complexity of TEC and a real-world example. We then outline a general overview of our approach, followed by an in-depth literature review of foundational and recent works in the domain of sentiment analysis and Transformer-based models. After this, the method implementation explains the dataset, preprocessing steps, and our hybrid BERT-RoBERTa architecture. Finally, the separate results and conclusion sections present empirical findings, confusion matrix analysis, and the implications of our model.

III. Literature Review

The authors in the paper by Sharma et al. [1] provide a detailed review of sentiment analysis, covering its tasks, applications, and the transition from classical machine learning to deep learning techniques. The paper highlights that Transformer-based models like BERT have significantly improved performance in TEC. They also emphasize that sarcasm and irony remain major challenges, motivating the need for more context-aware models.

The authors in the work of Dikbiyik et al. [2] introduce "BIMER," a bimodal system combining speech and text for emotion detection. Their findings show that bimodal systems outperform unimodal ones, and that data augmentation plays a crucial role in handling class imbalance. They also identify high design and computational complexity as future challenges.

The authors in the EEG-based study by Mohammed et al. [3] propose a physiological-signal-based approach using

This work was supported by Bennett University.

Roberts Similarity and PSO for feature selection. They demonstrate that EEG data captures nonlinear emotional variations effectively, although such systems require specialized hardware and are impractical for text-based TEC tasks.

The authors in the work of Joshi et al. [4] evaluate Transformer-based models, including BERT and RoBERTa, for six-class emotion detection. They find that Transformers outperform traditional models such as RNNs and SVMs. However, their dataset does not include complex emotions such as sarcasm or confusion.

The authors in the SemEval-focused paper by Paran et al. [5] explore cross-lingual and multilingual Transformer fine-tuning. They show that multilingual models can generalize to low-resource languages but still suffer a large performance gap compared to high-resource languages.

The authors in the research by Maryam et al. [6] explore the strength of RoBERTa for sentiment analysis on social media posts. They argue that RoBERTa is highly effective because it handles slang, informal text, and online conversational patterns, but note that social media's noisy nature remains a challenge.

The authors in the comprehensive review by Albladi et al. [7] focus specifically on sentiment analysis of Twitter data. Their findings verify that Transformer models are superior to older architectures and that preprocessing is vital for handling the high-volume, rapidly changing language on Twitter.

IV. Method Implementation

Our approach synthesizes these lessons: we build a hybrid model (inspired by [2], [7]) using BERT and RoBERTa (identified as top performers in [1], [4], [6]) and train it on a highly nuanced dataset, one that directly addresses key challenges like sarcasm [1], [4].

A. Dataset

We use the "Emotions Dataset" by boltuix [8], hosted on Hugging Face. This dataset contains 130k+ samples across 13 rich emotion labels such as Sarcasm, Confusion, Guilt, and Shame, enabling training on nuanced non-literal emotions.

B. Preprocessing

1) Label mapping of all 13 classes to integer values.
2) Train-test split: 90% training and 10% validation.
3) Dual tokenization using BERT and RoBERTa tokenizers at max length 128.

C. Hybrid BERT-RoBERTa Model

We use two pre-trained models in parallel. Each receives the same input, and their CLS embeddings (each 768-dim) are concatenated into a 1536-dim vector. This vector is fed into a linear classifier predicting 13 emotion classes.

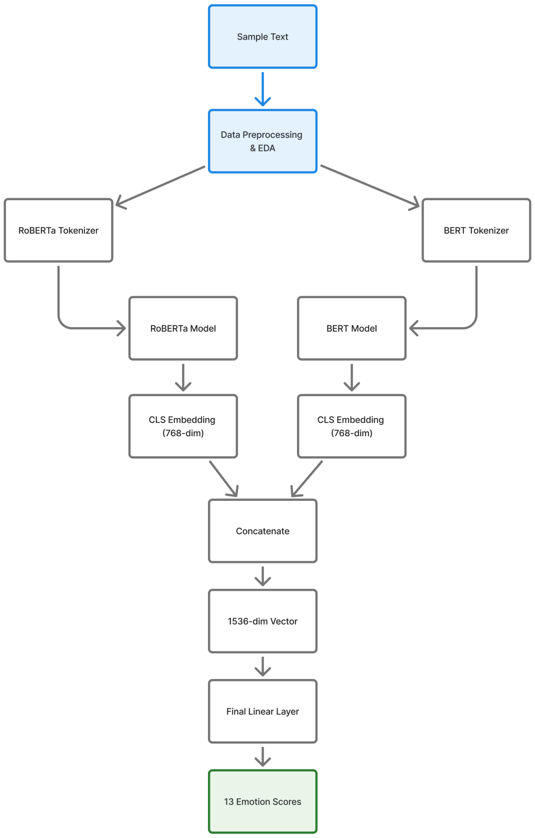

Fig. 1. The proposed hybrid BERT-RoBERTa architecture.

V. Results

Training was performed on an Nvidia RTX 4060 GPU for 5 epochs using the AdamW optimizer. The training loss decreased steadily from around 1.8 to 0.5, while the validation loss followed a similar trend, indicating effective learning without significant overfitting.

The confusion matrix in Fig. 2 shows strong diagonal activation, demonstrating high classification accuracy across most classes. Expected overlaps appear between related emotions, such as Sadness–Guilt or Anger–Disgust, but the model distinctly captures nuanced categories such as Sarcasm.

We also compare four models: Logistic Regression, BERT-only, RoBERTa-only, and our Hybrid model. As shown in Table I, the hybrid model outperformed all baselines with a validation accuracy of 69.93% and a weighted F1 score of 0.70.

VI. Conclusion

Our hybrid BERT-RoBERTa model is highly effective for subtle text emotion classification. Concatenating both feature spaces allowed for better detection of complex states such as sarcasm, confusion, and shame. With the 13-label dataset

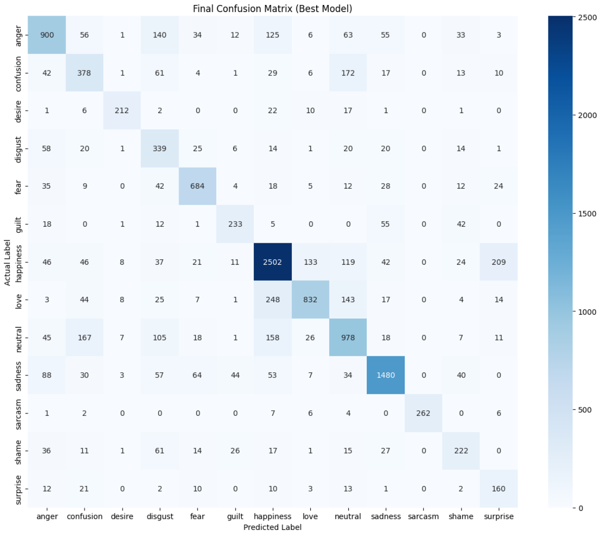

Fig. 2. Confusion matrix for the 13 emotion classes. High diagonal values indicate correct predictions.

TABLE I
COMPARISON TABLE OF FOUR MODELS (13-CLASS DATASET)

Model	**Validation Acc.**	**F1-Score (W.)**
Logistic Regression	41.2%	0.40
BERT-Only	65.3%	0.65
RoBERTa-Only	67.1%	0.67
Hybrid (Ours)	**69.93%**	**0.70**

enabling richer context understanding, our model directly addresses core challenges highlighted in foundational works [1], [4]. The results demonstrate that hybrid architectures combined with rich datasets pave a strong path forward for building more human-like emotion classifiers.

REFERENCES

[1] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, ”A review of sentiment analysis: tasks, applications, and deep learning techniques,” *International Journal of Data Science and Analytics*, 2024.

[2] E. Dikbiyik, O. Demir, and B. Dogan, ”BIMER: Design and Implementation of a Data Enhanced Bimodal Emotion Recognition System Augmentation Techniques,” *IEEE Access*, 2025.

[3] M. H. Mohammed, M. N. Kadhim, D. Al-Shammary and A. Ibaida, ”EEG-Based Emotion Detection Using Roberts Similarity and PSO Feature Selection,” *IEEE Access*, 2025.

[4] D. B. Joshi, A. A. Dhokai, and N. P. Agrawal, ”Emotion Detection Using Transformer Model With DeepLearning,” in *2024 5th International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2024.

[5] A. I. Paran, S. Aftahee, M. R. Hossan, J. Hossain, and M. M. Hoque, ”Fine-Tuning Deep Learning and Transformer-based Models for Emotion Detection in Multi-label Classification, Estimation of Intensity and Cross-Lingual Adaptation,” in *SemEval-2025*, 2025.

[6] Z. Maryam, et al., ”Sentiment Analysis on Social Media Posts Using Roberta,” *Journal of Computing & Biomedical Informatics*, 2025.

[7] A. Albladi, M. Islam, and C. Seals, ”Sentiment Analysis of Twitter Data Using NLP Models,” *IEEE Access*, 2025.

[8] Boltuix, ”Emotions Dataset,” Hugging Face, 2023. Available: https://huggingface.co/datasets/boltuix/emotions-dataset