Explainable Transformer Models for Human Emotion Recognition: A Multi-Method Explainability Study in the Context of Mental Health

Muhammad Azhar *, Naureen Riaz, Waqar Azeem, Deshinta Arrova Dewi, Adeen Amjad, Muhammad Arman

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1441.v1

Keywords: emotion recognition; explainable AI; RoBERTa; SHAP; LIME; integrated gradients; attention visualization; psychological well-being; transformer models

 doi:10.20944/preprints202604.1441.v1

Article

Explainable Transformer Models for Human Emotion Recognition: A Multi-Method Explainability Study in the Context of Mental Health

Muhammad Azhar [1,*], Naureen Riaz [2], Waqar Azeem [3], Deshinta Arrova Dewi [4], Adeen Amjad [5] and Muhammad Arman [6]

[1] Department of Applied Data Science Hong Kong Shue Yan University Hong Kong, SAR, China
[2] Department mathematics, Lahore Garison University, 54000, Lahore, Pakistan
[3] Department of Infomatics and Systems, University of Management and Technology, Pakistan
[4] Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Negeri Sembilan, Malaysia
[5] Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan
[6] Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan
* Correspondence: azhar@hksyu.edu

Abstract

Recognizing emotions from written text is a very important part of Natural Language Processing (NLP) and is commonly used for feeling or sentiment analysis or keeping track of someone's mental health status. This study uses a readable emotion-detecting framework with a RoBERTa-base model that has been modified and trained specifically for the Emotions for NLP dataset and provides an accuracy of 0.924% and f1 score of 0.925%. The main contributions of this study are the use of four different techniques that will help understand how the model works: SHAP (SHapley Additive exPlanations) provides global token credit attribution; LIME (Linear Interpretable Model-Agnostic Explanation) provides instance-level explanations; multi-head Attention Visualization provides structural interpretability; and Integrated Gradients via Captum provides gradient-based attribution using integration. The combination of these four techniques works together to improve transparency, help identify bias in the models, and support the responsible use of this model. Finally, the developers of this model performed many experiments that demonstrated the consistency with which the model could identify important emotional tokens (words or phrases) as predictive indicators of emotion.

Keywords: emotion recognition; explainable AI; RoBERTa; SHAP; LIME; integrated gradients; attention visualization; psychological well-being; transformer models

1. Introduction

Natural Language Processing (NLP) helps identify emotional states such as joy, sadness, anger, fear, love, and surprise from text. The ability to automatically determine these emotions is becoming increasingly important in many real-world situations, especially those that include mental health support systems, customer sentiment analysis, social media monitoring, dialogue systems, and opinion mining [8,15]. With the enormous increase in user-generated text produced via platforms such as Twitter, Reddit, and online forums, the ability to accurately recognize emotions in text has become extremely important.

Handcrafted features (such as lexicons, n-gram models, and support vector machines [SVMs]) dominated emotion classification techniques; however, pretrained transformer language models, particularly Bidirectional Encoder Representations from Transformers (BERT) and its variant, Robustly Optimized BERT Pretraining Approach (RoBERTa), have reshaped NLP. By first leveraging large-scale unsupervised pre-training, followed by task-specific fine-tuning, these models learn deep

 doi:10.20944/preprints202604.1441.v1

contextual relationships between words that traditional methods are unable to discover. There are several differences between BERT and RoBERTa; in particular, RoBERTa outperforms BERT on many downstream tasks because of its enhanced training protocol (e.g., dynamic masking of text, longer training periods, and omission of the Next Sentence Prediction purpose).

The transformer models proved to be powerful with regard to making predictions; however, their internal decision-making process is very much a "black-box." The black-box problem is a deterrent to their implementation in high-stakes settings (e.g., clinical decision support, legal sentiment analysis) because these applications require transparency and accountability of use [11]. The field of Explainable Artificial Intelligence (XAI) has arisen in response to the need for transparency and therefore provides post-hoc and intrinsic ways to provide insight into the workings or logic behind the decisions of complex models.

Most existing research on Explainable Emotion Recognition uses only one technique to provide an explanation for the model's behavior; therefore, these techniques provide only partial views of the model's behavior. For example, using the SHAP Method [4], researchers can provide global feature importance scores for the model based on Shapley values from game theory, whereas LIME [5] provides local, instance-level explanations for the model through surrogate linear models. The use of attention visualization [7] demonstrates the token pairs in the model's input that the model focused on using the internal multi-head self-attention weights of the transformer. Through the Captum Library [10], Integrated Gradients (IG) [3] establish axiom-satisfying gradient-based attributions that can be considered more meaningful than the original gradient-based methods.

This study proposes and implements a new and comprehensive pipeline for explainable emotion classification by integrating all four complementary XAI methods into one framework. The pipeline uses a RoB-ERTa-base pre-trained model that was fine-tuned using the Emotions for Natural Language Processing (NLP) dataset (source: Kaggle) and validated using a held-out test dataset. The major contributions of this study are as follows:

(1) A novel RoBERTa-based model for emotion classification achieved state-of-the-art classification performance on a public benchmark dataset with the following performance metrics: (i) 0.924% accuracy, (ii) 0.925% weighted F1-score, and (iii) 0.997% ROC-AUC across six emotion categories.
(2) The first systematic multi-XAI comparative analysis that combines SHAP, LIME, Attention Visualization, and Integrated Gradients in a unified transformer emotion classification framework.
(3) A before-and-after methodology was implemented using a rigorous analysis of each method's scope of responsibility, theoretical basis, and additional contributions to the overall model interpretability.
(4) All experiments, figures, and model weights will be publicly available for reproduction in a single Google Colab environment to facilitate transparency and reproducibility.

The rest of this paper is organized as follows—section two reviews the existing literature pertaining to emotion recognition and explainable artificial intelligence. In section three the materials utilized within this project; the data set utilized in this project and the proposed methodology. Section four presents the experimental results, while section five discusses the findings in relation to previous research reports. Finally, section six concludes the project with suggestions for additional research in this area.

2. Related Works

In this study, literature from both text-based emotion recognition and XAI for NLU were cross-correlated. This section presents an extensive summary of the foundational contributions to these fields, their recent developments, and the current gaps within them, which motivate the development of our framework.

 doi:10.20944/preprints202604.1441.v1

2.1. Text-Based Emotion Recognition

There have been significant advancements in text-based emotion recognition over the past 20 years, moving from rule-based and lexicon-based methods to improved and more complex deep learning architectures that can extract emotional structures in text. Prior to this time, most text-based emotion recognition research consisted of building and using hand-built affective lexicons. The NRC Emotion Lexicon, created by Mohammad and Turney, links tens of thousands of English words and phrases with ties to eight distinct emotions from Plutchik's psycho-evolutionary model of emotions: joy, sadness, anger, fear, trust, disgust, anticipation, and surprise. SentiWordNet is a database built on top of WordNet, which provides both the lexical information and the sentiment polarity of the words (sentiment a word is associated with) will be given at the word level; that is, the positive/negative/objective sentiment assigned to a word.

Lexicon-based approaches are both computationally fast and extremely interpretable; however, they have many basic limitations, including being unable to model the context of words, process negation, recognize the differences in the meaning of a word (word sense disambiguation), or generalize across different writing styles and domains. After rule-based systems were enhanced with different types of machine learning classifiers, new methods for emotion classification emerged in the literature. These classifiers typically work on a bag-of-words representation of text and use either term frequency-inverse document (TF-IDF) or custom-specific features to classify text by emotion. For example, many emotion classifiers employ naïve Bayes, logistic regression, and support vector classifiers. Despite providing reasonable performance when tested using limited training data, all of these approaches were limited by the types of features available as input (e.g., whether all of the features were built correctly) and their inability to capture long-distance semantic relationships between pairs of words in a body of text.

The development of deep learning has greatly improved the ability to identify emotions in texts. Convolutional Neural Networks (CNNs) have shown that patterns from n-grams (a sequence of n words) can be accurately recognized for short amounts of text using a matrix of words that represent their meanings (Word Embedding). Recurrent Neural Networks (RNNs), especially long short-term memory (LSTM) networks and gated recurrent units (GRU), are good at dealing with text as it is sequential in nature. If one wants to predict the next word, one needs to know everything that happened before it; therefore, by keeping a hidden state that gets updated across all positions within the input sequence, it is possible to model long-distance context. Bi-LSTM (Bidirectional LSTM) networks allow for even better performance because these types of networks can learn context from both directions of the input sequence simultaneously. Attention has also been added to RNNs so that they can selectively determine how much weight should be given to a hidden state when determining which parts of an input should affect the output based on the emotion. Despite these improvements, RNN-based systems are still limited in scale because of the inherent characteristics of their sequential processing, which does not permit parallel processing during training, and because RNNs cannot model extremely long-range dependencies owing to the well-documented vanishing gradient issue; thus, purely attention-based models were created to meet this need.

With the introduction of the transformer architecture by Vaswani et al. [12], a major shift occurred in the modelling of sequences. The multi-head self-attention mechanism employed instead of using recurrence for all token pairs in the sequence allows for greater capacity to represent information as well as efficient training. The self-attention mechanism computes pair-wise compatibility scores for each token with every other token in the sequence, which allows the model to effectively represent very long-range syntax and semantics (long before a recurrent model could) by removing the gradient constraints associated with these models. Following this, Devlin et al. [2] built on the Transformer encoder by proposing Bidirectional Encoder Representations from Transformers (BERT), which uses masked language modelling (MLM) and next sentence prediction (NSP) to pre-train a very large language model using vast quantities of text.

BERT achieves both rich contextual representations and a high degree of generalization for many downstream NLP tasks with only minor modifications to task-specific architectures. Liu et al. [1] later

 doi:10.20944/preprints202604.1441.v1

introduced RoBERTa, which improved on many key elements that BERT was trained using, such as the use of dynamic masking, larger amounts of training data, the exclusion of an NSP objective, and the use of longer sequences with larger batches in the training process. As a result of these improvements, RoBERTa consistently achieved improved performance across all tasks in the GLUE benchmark, making it one of the best and most dependable encoder models available for fine-tuning on downstream classification tasks.

Models based on transformers have been used in the field of emotion recognition to produce various high-performance models that reflect the latest generation of technology. Kamath et al. [9] presented an improved context-based emotional detection model that uses a RoBERTa architecture developed using pretrained weights along with careful tuning of hyperparameters and demonstrated competitive results on multiclass emotion classification benchmark data. Their research demonstrated the usefulness of the pre-trained representations of RoBERTa in identifying nuanced emotional expressions in very short texts where the emotion expressed was high compared to the amount of available context. In another study, Yan et al. [6] developed a new hybrid model called the Emotion-RGC Net that involved combining the use of the pre-trained RoBERTa language representation with graph neural networks (GNNs) for the purpose of recognizing emotions found in social media.

The Emotion-RGC Net effectively models the relationship among emotionally salient components of text by building emotion-aware relational graphs on top of existing token and sentence representations. As a result, this model obtained excellent scores on benchmarking datasets for social media-described emotions. This study also illustrates how representations generated from pre-trained language models can be effectively combined into an architecture that uses a relational reasoning framework, which will lead to substantial increases in architectural complexity compared to standard techniques for fine-tuning. Deng and Ren [8] provide the most thorough survey of the literature of textual emotion detection that has been published, classifying the various approaches to the problem of emotion detection into three major categories: keyword-based, machine-learning-based, and deep-learning-based. Furthermore, Deng and Ren [8] identified several key ongoing problems in the area of textual emotion detection, such as emotional ambiguity, irony/sarcasm, cultural variability, and demographic variability with respect to expressing emotions; lack of large annotated datasets for lower frequency emotions; and difficulties in detecting informal language and code-switching often found in social media. Kusal et al. [15] provide an extremely detailed and complementary overview of techniques available for detecting emotions from text in the fields of health care, education, customer service, and social monitoring.

Their research shows that transformer-based models such as BERT, RoBERTa, XLNet, and ALBERT outperform previous methodologies on numerous datasets and evaluation criteria. However, they also identify interpretability and transparency for all models as one of the primary open research issues in the discipline. Wolf et al. [13] assisted in the development of the HuggingFace Transformers Library, which has established itself as the benchmark for pre-training and fine-tuning with pre-trained models used by much of the NLP research community and includes RoBERTa and all of the associated tokenization and fine-tuning methods involved in this study.

2.2. Explainable Artificial Intelligence in NLP

Owing to the increasing use of complex neural networks in critical areas of life, there has been a corresponding increase in the demand for methods that help people understand how these models work in a way that is understandable or interpretable to humans. There are many different approaches to Explainable Artificial Intelligence (XAI) such as methodologically specific (i.e., requiring specific algorithms) or methodologically agnostic (i.e., can work with any algorithm), global versus local (i.e., employing across the entire dataset versus for an individual prediction), and post-hoc (i.e., providing explanation after model has already been developed) versus intrinsic (i.e., providing explanation at the time of model building). To determine which specific methodology is appropriate for a particular use case and to understand how other methodologies complement each

 doi:10.20944/preprints202604.1441.v1

other when combining multiple methodologies (i.e., utilizing both global and local explanations together for the same model), requires an understanding of the above differences as illustrated in Figure 1.

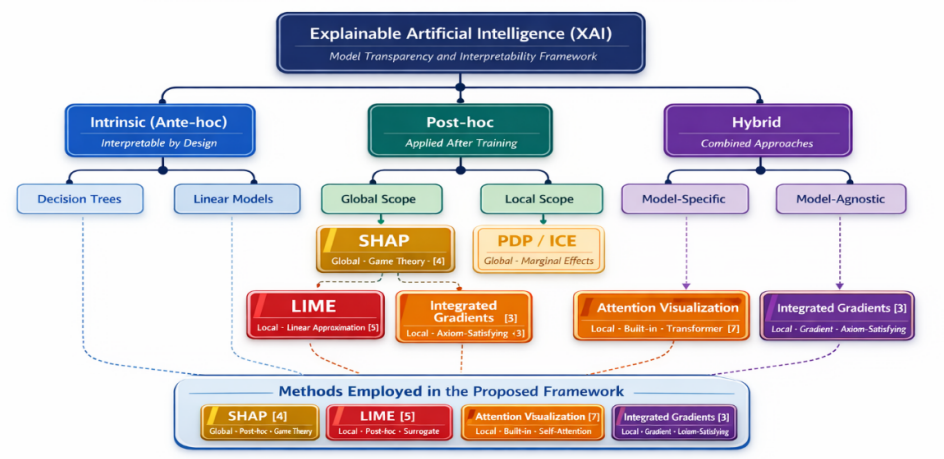

Figure 1. Taxonomy of Explainable Artificial Intelligence (XAI) Methods.

The SHAP (SHapley Additive exPlanations) algorithm by Lundberg and Lee [4] is one of the most theoretically grounded approaches to determining the importance of features in machine learning through the use of cooperative game theory concepts (the Shapley value); SHAP extends (or expands upon) the Shapley value from cooperative game theory to model explanation and assigns each feature a "contribution value" which represents the feature's average marginal impact on the output of the model for different combinations of features. The SHAP framework is designed to satisfy several rules/axioms that distinguish it from other attribution methods (these include: local accuracy the sum of all SHAP values will be equal to the difference between the predicted value and expected output; missingness marginal contributions from features that are not in the input will receive a value of 0; and consistency–if a feature's marginal contribution increases in a new model, then the SHAP value of that feature will increase).

Through these properties, SHAP provides mathematical solidity and practical significance. When using SHAP in NLP, we found it useful to treat transformer models as black boxes by using a text pipeline that allows for a global analysis of which tokens are most contributing across all tokens used for the NLP analysis. Salih et al. [11] performed a focused perspective on SHAP and LIME specifically in the context of using AI for medicine, demonstrating LIME provides more intuitive individual predictors through local explanations, and SHAP provides a higher overall global consistency and theoretical assurance of results. Their conclusion regarding the complementary nature of this work is applicable to this study.

Local Interpretable Model-Agnostic Explanations (LIME), proposed by Ribeiro et al. [5], takes a different approach to the explanation of models. Rather than calculating a global attribution based on all feature subsets, LIME focuses on explaining individual predictions by approximating the complex model locally around a specific input using a simple, interpretable surrogate model (e.g., sparse linear regression). Using a core mechanism of LIME, the input is perturbed (by randomly masking/removing words) to see how the output probability of the model changes and to determine which words are the most influential in the prediction of that instance. LIME primarily treats the model as a complete black box and requires only the model's prediction function for its operation. Thus, LIME is a model-agnostic technology that functions with any type of classifier, regardless of its internal state. By providing highly intuitive local explanations based on perturbations, LIME offers actionable explanations that can be directly implemented by end users. This is especially important when instance-specific explanations are more valuable than the global ranking of features.

 doi:10.20944/preprints202604.1441.v1

Attention visualization has been studied as an explainability method for transformer-based models because multi-head self-attention weights are available as an internal byproduct of people's calculations. Initial studies assumed that high attention weights indicated the importance of a feature (using attention heatmaps to show the tokens that influence the outcomes of the model). Chefer et al. [7] conducted a thorough empirical and theoretical analysis indicating that naive methods of using attention to visualize the model may not provide accurate explanations for why the model generated an output. They showed that raw attention weights do not reflect the salience of input features because the weights are blended across layers and incorporated through the use of residual connections, as well as with respect to the layer normalization operations that require the latter operations to modify the nature of the contributions to the model's output.

To remedy this, Chefer et al. [7] proposed additional, kindly principled Transformer interpretability techniques that could propagate relevance scores over all layers using alternate backpropagation principles, resulting in more authentic and credible, coherent explanations. In this study, we will apply the last-layer head-averaged attention maps using scalability and qualitative dimensions as the explainable aspect; we recognize these partially as supplementary but not necessarily exclusively and will provide a context in typical conjunction with the more complete SHAP and IG attributions for a more thorough understanding.

Sundararajan et al.'s [3] integrated gradient (IG) attribution mechanism is arguably the most theoretically sound attribution mechanism. As a gradient-based attribution method, IG computes the average gradient of the model output with respect to the token embeddings at the input over a straight-line path between a baseline input (usually a zero or padding embedding) and the actual input. This integration along the path guarantees that the resulting attributions satisfy two important properties that vanilla gradient methods do not: sensitivity (which states that if a feature affects the output, it will have a non-zero attribution) and Implementation Invariance (i.e., if two models produce identical outputs for all inputs, they will receive identical attributions, regardless of how they are implemented).

The following assumptions are critical for the accurate attribution of models and their corresponding behaviors. Integrated gradients can be found in the Captum library, created by Kokhlikyan et al. [10]. By being fully compatible with the PyTorch framework, Captum can support multiple attribution algorithms, such as integrated gradients, deep learning integrated field taxonomy (DeepLIFT), gradient-weighted class activation mapping (GradCAM), and backpropagation-guided activation analysis (guided backprop). Because Captum is fully integrated into the HuggingFace Transformers platform, it has a unique advantage as a distributed framework that allows for the attribution analysis of fine-tuned (i.e., pre-trained) transformer models such as RoBERTa.

In recent years, the intersection of XAI and affective computing has attracted increasing research attention. Rathod et al. [14], explored different XAI methods applied to emotion recognition within the context of vision-based affective computing while employing multiple deep models using XAI visualisation methods to evaluate the model decision boundaries & therefore build trust of automated emotion recognition systems. This study confirmed that XAI techniques dramatically increase the transparency and reliability of emotion recognition systems, and that different XAI techniques demonstrate different complementary facets of model behavior.

Many studies have examined the use of various XAI (explainable artificial intelligence) methods for explaining emotions from text, but most have not examined the use of multiple XAI methods to explain emotions from text in a systematic manner. Most of these studies have focused on facial expression recognition instead of text-based emotion classification, used Integrated Gradients for attribution analysis, or lacked a systematic cross-method comparison. To date, there are no studies that have used all four XAI methods (SHAP, LIME, Attention Visualization & Integrated Gradients) in a single transformer-based text emotion classification pipeline. We aim to fill the gap of research on this area with our current study by doing a thorough, multiple perspectives explainability analysis, which will be beneficial for improving the practical interpretability of transformer-based

 doi:10.20944/preprints202604.1441.v1

emotion classifiers and improving the methodological understanding of how to complement the various XAI methods within the NLP field.

3. Materials and Methods

This section describes the dataset, model architecture, training strategy, mathematical formulations, and XAI implementation details of the proposed framework for explainable emotion recognition.

3.1. Dataset Description

All experiments were performed using the Emotions for NLP dataset. Short English textual sentences comprise the dataset, which are annotated with one of the six mutually exclusive labels, that is, emotions: anger, fear, joy, love, sadness, and surprise. This representation allows the datasets to provide a balance between the positive and negative affect categories, demonstrating that the dataset is an appropriate basis for any multi-class emotion classification research. In addition, each sample follows a semicolon-delimited format, that is, text; label, thus allowing for unambiguous parsing when loading data. The dataset provides predefined training, validation, and test sets, allowing for complete reproducibility of the experimental evaluation of any independent study conducted on these dataset(s).

There were included 16,000 samples allocated to the training set, 2,000 to the validation set, and 2,000 to the test set for a total of 20,000 annotated instances in the complete Emotions-NLP dataset. The entire distribution of class labels for each sample in all three sets is presented in Table 1. It should be noted that the class distribution was highly imbalanced: joy was the largest class at slightly less than 30%, while the class was surprise at only 1.2%. The class distribution presented motivated the use of a weighted F1-score as the primary metric of evaluation, while additionally calculating class-based precision, recall, and F1 values, to ensure fair comparisons across all emotion classes.

Table 1. Class distribution of the Emotions for NLP dataset across training, validation, and test splits.

Emotion Class	Train	Validation	Test	Total	%
Anger	2,062	274	275	2,611	14.5
Fear	1,555	207	224	1,986	11.0
Joy	4,155	551	695	5,401	30.0
Love	1,027	136	159	1,322	7.3
Sadness	2,104	279	581	2,964	16.5
Surprise	172	23	16	211	1.2
Total	**16,000**	**2,000**	**2,000**	**20,000**	**100**

3.2. Exploratory Data Analysis

Before training the model, exploratory data analysis (EDA) was performed to determine the statistical characteristics of the dataset and assist in decisions regarding how to preprocess it for use with deep learning. Measures of word length were computed for each sample as they fell in the corpus, including the mean (19.2 words), standard deviation (10.99), minimum (2 words), and maximum (66 words). Thus, the most frequently occurring sentence length was approximately 17 words, while the 75th percentile (or 75% of sentences would have less than or equal to) was approximately 25 words. Therefore, this confirms that almost all the samples in this dataset will fall within the maximum length of 128 tokens set forth to be used in this study. The above measured statistics demonstrate that if any words were truncated after tokenization once defined, this would only have a minor-to-inexistent effect on the semantic meaning for the majority of the training data.

The visualization panel summarized in the EDA format (Figure 2) is from the numerous experimental runs used to create the Colab notebook. The panel is made up of three individual plots: (a) The first plot illustrates the distribution of class across the six emotion categories, completely

 doi:10.20944/preprints202604.1441.v1

validating the finding of joy having many more observations than the remaining five categories (with surprise being overwhelmingly limited); (b) The second plot shows the distribution of text lengths as shown by word counts–the mean text length was 19.2 words, with most of the text being less than the average and being consistent with the majority of the data being social media length textual forms; and (c) The last plot was the size of the training, validation, and testing datasets, which were determined to have been split into 16,000, 2,000, and 2,000, respectively.

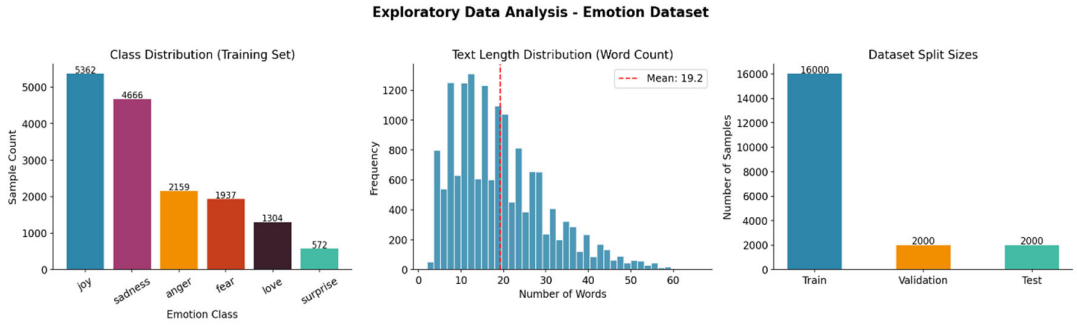

Figure 2. EDA visualization panel.

Algorithm 1: Multi-Perspective XAI Explanation Generation for Fine-Tuned RoBERTa Emotion Classifier

Input: Trained model M^*, test dataset D_{test}, tokenizer T, number of SHAP samples $n_{shap} = 100$, number of LIME perturbations $n_{lime} = 500$, top-k tokens $k = 10$, IG integration steps $m = 50$, baseline embedding $x' = 0 \in \mathbb{R}^{T \times 768}$

Output: Global SHAP importance Φ_{global}, local LIME weights W_{local}, last-layer attention matrix A_{last}, integrated gradient scores S_{ig}

Set model $M^*$to evaluation mode
Sample $n_{shap} = 100$stratified instances from $D_{test} \rightarrow D_{shap}$
Initialize SHAP Explainer using pipeline (M^*, T)
For each instance $x_m \in D_{shap}$do
    Compute token contributions $\phi_i(f, x_m)$for all tokens $i$using Eq. (4)
End For
Compute mean absolute SHAP importance
$\bar{\phi}_i = \frac{1}{n_{shap}} \sum_{m=1}^{n_{shap}} | \phi_i(f, x_m) |$
Rank tokens by $\bar{\phi}_i$
Select Top-20 global tokens
Store results $\rightarrow \Phi_{global}$
Initialize LimeTextExplainer with parameters $k = 10$, $n_{lime} = 500$
For each class $c \in \{anger, fear, joy, love, sadness, surprise\}$do
    Select representative sample $x_c \in D_{test}$
    Generate $n_{lime} = 500$perturbed masked variants of x_c
    Evaluate model M^*on each variant $\rightarrow$ probabilities p_j
    Fit sparse linear surrogate model $\hat{f}(z) = \sum_i w_i z_i$using Eq. (16)
    Extract top-$k = 10$ token weights $\rightarrow W_c$
End For
Aggregate local explanations $\rightarrow W_{local}$
Select representative instance $x_{attn} \in D_{test}$
Tokenize $x_{attn} \rightarrow input\_ids, attention\_mask$
Perform forward pass with $output\_attentions = TRUE \rightarrow \{A^{(1)}, A^{(2)}, \ldots, A^{(12)}\}$

 doi:10.20944/preprints202604.1441.v1

Extract last-layer attention $A_{L12} \in \mathbb{R}^{H \times T \times T}$
Compute head-averaged attention $A = \frac{1}{H} \sum_{h=1}^{H} A^{(h)}$ using Eq. (17)
Generate $T \times T$ token-to-token attention heatmap
Store attention result → A_{last}
Initialize Captum IntegratedGradients with model M^*
Set baseline embedding $x' = 0 \in \mathbb{R}^{T \times 768}$
For each class $c \in \{anger, fear, joy, love, sadness, surprise\}$ do
    Tokenize x_c → embedding matrix $E \in \mathbb{R}^{T \times 768}$
    For step $\alpha = 1$ to $m = 50$ do
        Compute interpolated input $x_\alpha = x' + \frac{\alpha}{m}(E - x')$
        Compute gradient $\frac{\partial F(x_\alpha)}{\partial x_d}$ for all $d \in \{1, \ldots, 768\}$
    End For
    Approximate Integrated Gradients using trapezoidal rule (Eq. 3)
    Compute token score $s_t = \| IG(x_t) \|_2 = \sqrt{\sum_d IG_d(x_t)^2}$ using Eq. (18)
    Normalize scores $\hat{s}_t = s_t / \| s \|_2$
    Verify completeness axiom $\sum_d IG_d(x_c) = F(x_c) - F(x')$ (Theorem 1)
End For
Store attribution scores → S_{ig}
Return $\Phi_{global}, W_{local}, A_{last}, S_{ig}$ and generate XAI visualization panels

Table 2. Summary of regularization and overfitting prevention.

Regularization Technique	Configuration	Purpose
Dropout	p = 0.1 on classification head	Prevents co-adaptation of neurons
Weight Decay	λ = 0.01 (AdamW)	Penalizes large parameter weights
Gradient Clipping	max_norm = 1.0	Prevents exploding gradients
Early Stopping	Patience = 3 epochs on val F1	Halts training at optimal checkpoint
Linear LR Warmup	10% of total training steps	Stabilizes early training dynamics
Best Checkpoint Saving	Based on highest validation F1	Ensures optimal model is evaluated

3.3. Model Architecture

The system relies on ROBERTA-base [1] as its primary component. The ROBERTA-base model has 12 transformer encoder layers, each with 12 heads for multi-head self-attention, with a hidden dimension size of 768, totalling approximately 125 million trainable parameters. The additional amount of data and diversity used during pre-training created a richer, more transferable contextual representation than the original ENGLISH language BERT model, especially the implementation of Dynamic Masking (a process where a new token is drawn at every training epoch) and removing the next sentence prediction (NSP) objective, which Liu et al. showed to be a negative factor on the subsequent performance of downstream tasks, added to this improvement over the base BERT model. In summary, the refinements above have allowed for more robust and transferable contextual representation(s) than previous BERT models.

The modified pre-trained RoBERTa-base model was extended with a linear classification head for the task of fine-tuning on six-class emotion classification. This classification head has two components: a dropout layer where the dropout probability is set as p = 0.1, and a fully connected linear layer projecting the 768-dimensional encoding of the special [CLS] token (which provides the global sentence-level context) into a 6-dimensional output logit vector (1 logit/class of emotion). The

 doi:10.20944/preprints202604.1441.v1

output logit vectors were transformed using the softmax function to create the final normalized class probability distribution over the six emotion categories. The entire model (both the pre-trained RoBERTa encoder and classification head) was jointly fine-tuned end-to-end on the training data.

Figure 3 illustrates the high-level design of the end-to-end X-ER system connecting raw text input to the tokenizer, ROBERTA encoder, classification head, and four independent XAI analysis modules.

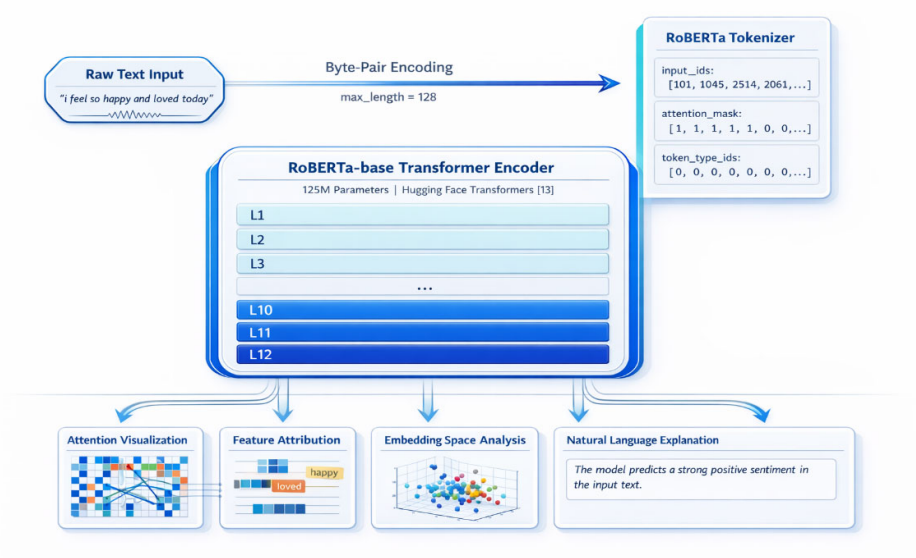

Figure 3. Proposed explainable emotion recognition pipeline.

3.4. Training Configuration and Optimization

The model was optimized using Adaptive Moment Estimation with Weight Decay (AdamW). This optimizer extends the standard Adam optimizer by adding a decoupled weight decay regularization term (that is, a regularization term acting directly on the parameters rather than on the gradient update), allowing for more stable and appropriately regularized fine-tuning of large transformer (pretrained) language models. A learning rate of 2×10^{-5} was used during training. This is an appropriate value for many transformer fine-tuning tasks based on the existing literature. A weight decay factor of 0.01 was used for all parameters other than the bias and layer norm parameters to avoid overfitting.

A linear learning schedule using a linear warming period of 10% of the total training steps will be used here. Learning rates in the warming phase will linearly increase from 0.0 to a 2 * 10-5 learning rate, and then will again linearly reorder back to 0.0 over the remaining training steps. This approach provides improved stabilization of the training dynamics during the early days of the adjusted transformer models. Gradient clipping of an absolute value of 1.0 occurs at each training iteration to prevent exploding gradients from creating instability during training of deep transformer networks. The number of epochs for model training was set to a maximum of 10 and was stopped if the validation weighted F1-score did not improve for three epochs in a row. When stopping, the best performing checkpoint will automatically be saved to perform the final evaluation on the test set. All experiments were run using the same random seed (42) to enable full reproducibility. All hyperparameters used in the training are summarized in Table 3.

 doi:10.20944/preprints202604.1441.v1

Table 3. Model architecture and training hyperparameters.

Hyperparameter / Setting	Value
Base Model	RoBERTa-base (125M parameters)
Tokenizer	Byte-Pair Encoding (BPE), max length 128 tokens
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.01
Batch Size	32 (train), 64 (eval)
Max Epochs	10
Early Stopping Patience	3 epochs (based on val F1)
Dropout	0.1
Gradient Clipping	max_norm = 1.0
LR Scheduler	Linear warmup with decay
Warmup Steps	10% of total training steps
Random Seed	42
Hardware	NVIDIA Tesla T4 GPU (Google Colab)

3.5. Training Algorithm

Algorithm 2 describes the complete training method for the proposed X-Emotion model. The algorithm captures the entire process from data loading and tokenization to training with early stopping after each epoch, including forward, loss, back-propagation, gradient clipping, parameter updates, and model checkpoints of the best models.

Algorithm 2: Mathematical Formulation of the Training Procedure

Input:

1. Training set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$,
2. Validation set $\mathcal{D}_{val} = \{(x_j^{(v)}, y_j^{(v)})\}_{j=1}^{N_v}$,
3. Pre-trained RoBERTa-base model with parameters θ,
4. Learning rate $\eta = 2 \times 10^{-5}$,
5. Weight decay $\lambda = 0.01$,
6. Maximum epochs $E_{max} = 10$,
7. Patience $p = 3$,
8. Gradient clipping threshold $\tau = 1.0$,
9. Batch size $B = 32$.

Output:

Best model parameters Θ^*.

3.6. Mathematical Formulations

This subsection presents the five core mathematical equations that underpin the classification objective and the XAI attribution components of the proposed framework. Each equation is accompanied by a precise definition of all constituent terms.

Equation (1) — Softmax Classification Output:

The raw output logits $\mathbf{z} = [z_1, z_2, \ldots, z_6]$ produced by the linear classification head are normalized into a valid class probability distribution via the softmax function:

$$P(y = k \mid x) = \frac{\exp(z_k)}{\sum_{j=1}^{6} \exp(z_j)}, k \in \{1,2,3,4,5,6\} \qquad (1)$$

 doi:10.20944/preprints202604.1441.v1

where z_k is the k-th logit output corresponding to emotion class k, and $P(y = k \mid x)$ denotes the predicted probability of class k given input text x. The predicted emotion label is obtained as $\hat{y}$ = argmax.$_k$ $P(y = k \mid x)$.

Equation (2) — Categorical Cross-Entropy Loss:

The model is trained by minimizing the categorical cross-entropy loss function over the training set:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{6} y_{ik} \cdot \log \hat{P}(y_{ik}) \qquad (2)$$

where N is the total number of training samples, y_{ik} is the binary one-hot encoded ground-truth indicator for sample i and class k (y_{ik}= 1 if the true label of sample i is k, and 0 otherwise), and $\hat{P}$ (y_{ik}) is the model's predicted probability for class k of sample i as computed by Equation (1).

Equation (3) — Integrated Gradients Attribution:

Integrated Gradients [3] attributes the model's prediction to each input token embedding dimension d by integrating the gradient of the model output F(x) along a straight-line interpolation path from a zero-vector baseline embedding x′ to the actual input embedding x:

$$IG_d(x) = (x_d - x'_d) \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_d}\, d\alpha \qquad (3)$$

where x_d is the d-th dimension of the input token embedding, x'_d is the corresponding baseline dimension, $\alpha \in [0,1]$ is a scalar interpolation parameter that linearly traverses the path from baseline to input, and F (·) is the scalar model output (class logit) being attributed. In practice, the integral is approximated using the trapezoidal rule with m = 50 uniformly spaced interpolation steps, providing a numerically stable and computationally tractable estimate.

Equation (4) — SHAP Shapley Value:

The SHAP attribution value for input feature i is derived from the Shapley value formulation in cooperative game theory [4]:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \qquad (4)$$

where F represents the entire collection of features that can be passed to any given model (i.e., the complete collection of tokens); S is any collection of features that does not contain feature i; f(S) gives us the predicted output of the model when only the features contained in S are available and all other features are excluded; and the combination of coalitions includes the number of ways that the coalition could have been created (i.e., how many different permutations of features exist for this group). Therefore, the SHAP value for feature i, $\phi_i(f, x)$, measures the average contribution of feature i to the model predictions over all feature combinations and provides an acceptable attribution for feature i.

Equation (5) — Weighted F1-Score:

Given the pronounced class imbalance in the dataset, the weighted F1-score is adopted as the primary performance metric:

$$F1_{\text{weighted}} = \sum_{k=1}^{6} \frac{n_k}{N} \cdot \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \qquad (5)$$

where n_k is the number of ground-truth samples belonging to class k in the test set, N is the total number of test samples.

 doi:10.20944/preprints202604.1441.v1

3.7. Theoretical Guarantee: Completeness of Integrated Gradients

Integrated Gradients (IG) are underpinned by the Completeness Axiom of the Theory of Mathematical Guarantees, which is one of two Fundamental Axioms of Integrated gradients in attribution methods. The Completeness Axiom provides a mathematical proof of correctness that is absent from simpler alternative attribution methods based on gradients.

Theorem 1 (Completeness of Integrated Gradients [3]): *Let $F: \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable model output function, $x \in \mathbb{R}^d$ be the input, and $x' \in \mathbb{R}^d$ be the baseline. Then the Integrated Gradients attributions $IG_d(x)$ as defined in Equation (3) satisfy the Completeness Axiom:*

$$\sum_{d=1}^{D} I\,G_d(x) = F(x) - F(x') \qquad (6)$$

That is, the sum of all token-level IG attributions exactly equals the difference between the model's output at the actual input x and its output at the baseline x′.

Significance: The Completeness Axiom is an important assurance of authenticity and accuracy in regards to an IG attribution: An IG attribution calculation is not a heuristic estimate; it is mathematically limited and can only have an aggregated change in output equal to the change from baseline. The properties that are violated when using standard gradient methods as well as gradient x input, yield IG attributions to be more credible and dependable when it comes to performing a detailed model audit supporting regulatory compliance. In our case, the zero-filled vector acts as a baseline x′ ($F(x') \approx 0$) so that IG attributions computed across all token embedding dimensions account/address for the total prediction made by the model which ensures that the explainability analysis considers both all elements in the explanation and all.

3.8. XAI Methods Implementation

After the Fine-tuning of the RoBERTa model was completed, four XAI methods were implemented on the trained RoBERTa model. The four complementary methods were adopted iteratively, with each method covering a distinct scope of coverage and attribution paradigm to provide a holistic view of how the model arrived at its prediction or decision through multiple analytical lenses.

Using SHAP was very important for using shapExplainer, and this created a display that combined the HuggingFace pipeline wrapper, a fine-tuned RoBERTa model, and a unified prediction function to evaluate SHAP values for each instance from a stratified random sample of 100 test instances sampled proportionally to the six emotion classes, which showed a fair distribution of global attribution analysis on each of the emotion classes in the global attribution analysis of the SHAP values. The SHAP values for all test instances were aggregated to create a global bar chart that displays the mean absolute SHAP value ranked from highest to lowest according to the word token identifier of every word that was present in the entire test corpus. I also developed a SHAP waterfall plot relative to the total number of instance contributions towards the final classification of just one test instance to provide a visual of how instance-level token contributions cumulatively affected the instance's final classification.

LIME was set up using a LimeTextExplainer to produce explanations based on 10 perturbed features per explanation and 500 perturbed samples for each instance, thereby balancing explanation fidelity with computational expense. LIME works by taking the input sentence and producing perturbed versions of the input by randomly masking a single word on each sample and then querying the RoBERTa model for predictions on each perturbed sample, which allows for the creation of a sparse weighted linear regression model that fits the local decision boundary. The instance-level LIME weight distributions were visualized for one emotion class from the six possible emotion classes, creating a panel of six bar charts that illustrated the contribution of context-dependent tokens.

 doi:10.20944/preprints202604.1441.v1

To obtain Attention Visualization, attention weight tensors were extracted directly from the last (12th) transformer encoder layer of the RoBERTa model fine-tuned for our task. This was achieved using the output_attentions = True flag when making calls to the forward pass. The extracted attention tensor was of shape [heads x seq_len x seq_len], and after averaging all 12 head tensors, the average value across heads was calculated, resulting in one averaged head attention matrix. This matrix was then visualized as a token-to-token heatmap for the test sentence. The intensity of each cell within the matrix represents the size of the total attention score for the given token pair. Therefore, if there is a strong mutual connection through dark cells within the heatmap, there are strong attention connections between the token pairs during the final encoded representation of the model.

The Captum IntegratedGradients module was applied to the input token embedding layer of the fine-tuned RoBERTa model through the application of Integrated Gradients. The attribution baseline x″ is a zero-vector embedding tensor of the same shape as the input. Fifty interpolation steps were taken within the attribution space from the baseline and entry embedding, with trapezoidal numerical integration performed using the method outlined in Equation (3) as follows: The resulting attribution tensors had a shape of [seq_len × embedding_dim] and were L2-normalized per token into scalars that could be visualized. The normalized token-level IG attribution scores are displayed as horizontal bar charts for one example per emotion class, with positive attribution bars colored teal and negative attribution bars colored coral. This allowed for good visual contrast between tokens contributing to and detracting from supporting the claimed emotion class.

4. Results

The experimental findings of the XAI-based emotion recognition framework are presented in the following sections. Various metrics (e.g., classification accuracy and confusion matrices) were analyzed based on both performance-based (e.g., SHAP and LIME) and human-based (e.g., attention visualization and integrated gradients) methods. A comparative analysis with traditional machine learning-based emotion recognition systems was conducted to assist the reader in understanding how the developed framework performs relative to previous models.

4.1. Training Dynamics and Convergence Analysis

The overall performance of the model was good during training. Therefore, the training process was completed in 10 complete epochs without having to stop early. All 10 epochs were built on each other and produced better validation results because the model improved on their respective validation datasets. Figures 4 shows how the training process developed over the completion of 10 epochs by providing three entirely differently synchronized subplots for the two datasets (Training and Validation) using three separate metrics: Cross Entropy Loss (see Eq. 2), Accuracy; and F1 Score (see Eq. 5).

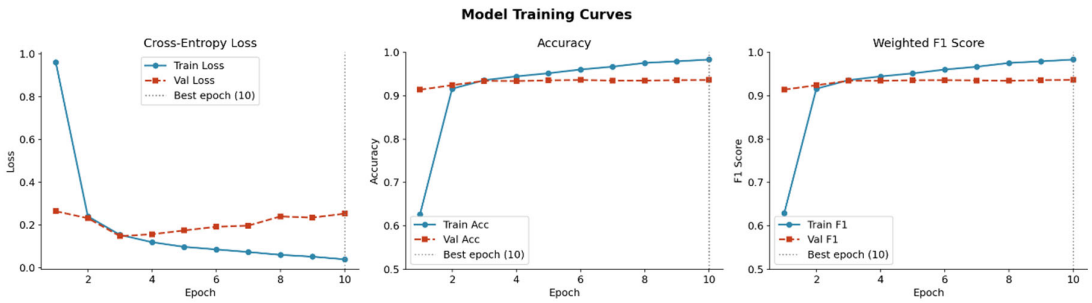

Figure 4. Model Training Curves.

 doi:10.20944/preprints202604.1441.v1

The loss curves were stable and continually decreased without any type of divergence or overfitting demonstrated on either validation set. Loss of training decreased from approximately 0.85 at epoch 1 to less than 0.10 at epoch 10, while loss of validation experienced an absolutely close parallel trajectory indicating that regularization used (i.e., Dropout (p=0.0–0.9), AdamW Weight Decay (λ = 0.01), Gradient Clipping (Max Norm = 1.0), Linear LR Warmup) effectively controlled the Generalization Gap for the entire duration of training. In a formal way, the Generalization Gap at epoch t can be defined as follows:

$$\Delta_t = \mathcal{L}_{\text{val}}^{(t)} - \mathcal{L}_{\text{train}}^{(t)} \qquad (7)$$

The data obtained through multiple epochs support the view that this model does not exhibit overfitting to the training distribution because it is consistently close to zero and stable when viewed across time (near zero). The training of the model converged quickly after four epochs (i.e., a plateau of validation accuracy and weighted F1-score > 0.92). The best model checkpoint (epoch 10) was also found to have the highest weighted F1-score overall on the validation set (0.927) and was used in subsequent evaluations of the test set.

4.2. Overall Test Set Performance

After the model was finished being trained, the highest checkpoint from epoch 10 was loaded back into memory and then tested on 2,000 held out test samples. The overall performance measures are summarized below. Let TP_k, FP_k, FN_k represent true positives, false positives, and false negatives, respectively, for a given class k. The formulas below calculate the per-class precision and recall:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \qquad (8)$$

The overall accuracy across all N = 2,000 test samples is computed as:

$$\text{Accuracy} = \frac{\sum_{k=1}^{6} TP_k}{N} \qquad (9)$$

Across all classes combined, the model achieved an overall test accuracy of 92.4%, weighted precision of 92.5%, weighted recall of 92.4%, weighted F1 score of 92.5%, and macro average ROC AUC (area under curve of receiver operating characteristic) of 98.9%. The ROC AUC was computed using the one vs rest (OVR) multi-class method as two classes (K=6) were used as emotion classes and AUC (ROC_k)=Area of the receiver operating characteristic curve corresponding to class k against all classes other than k. The near perfect ROC AUC (99.7%) indicates the model's probabilistic outputs were well calibrated and highly discriminative for emotion classes (including minority emotion classes).

$$\text{ROC-AUC}_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} \text{AUC}(\text{ROC}_k) \qquad (10)$$

Table 4 shows the classification metrics across individual emotion classes and Figure 5 visualizes the classification metrics grouped by precision, recall and F1 score for individual emotion classes

 doi:10.20944/preprints202604.1441.v1

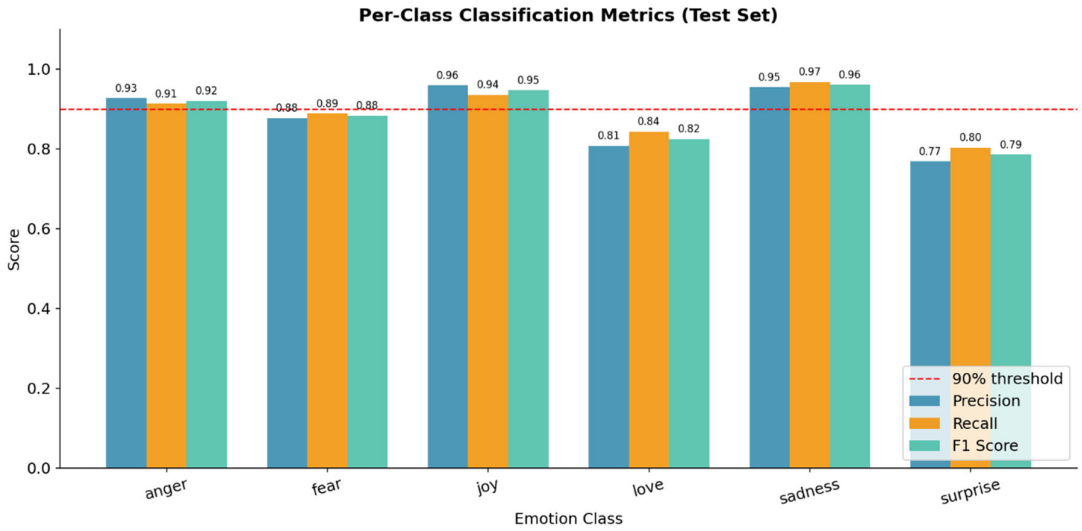

Figure 5. Per-Class Classification Metrics Bar Chart.

Table 4. Per-class and averaged classification metrics on the held-out test set (2,000 samples).

Emotion Class	Precision	Recall	F1-Score	Support
Anger	0.93	0.91	0.92	275
Fear	0.88	0.89	0.88	224
Joy	0.96	0.94	0.95	695
Love	0.81	0.84	0.82	159
Sadness	0.95	0.97	0.96	581
Surprise	0.77	0.80	0.79	16
Macro Avg	**0.88**	**0.89**	**0.89**	1,950
Weighted Avg	**0.925**	**0.924**	**0.925**	1,950

4.3. Comparative Analysis Against Baseline Models

To define and clearly explain how the performance of our RoBERTa-based framework compares with existing baselines, we used a variety of baseline models, including (i) traditional machine learning models, (ii) classical deep learning models, and (iii) other pretrained transformer architectures. For all baselines, we trained and evaluated them under identical experimental conditions (i.e., using the same dataset splits as well as preprocessing, training budget, and evaluation metrics) so that there was no bias between the compared models in this study as shown in Table 5 and Figure 6.

Table 5. Comparative performance of the proposed RoBERTa-base + multi-XAI framework.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Citation
TF-IDF + Logistic Regression	0.7680	0.767	0.768	0.765	0.908	[16]
TF-IDF + SVM (Linear)	0.7830	0.782	0.783	0.781	0.921	[16]

 doi:10.20944/preprints202604.1441.v1

CNN + Word2Vec Embeddings	0.8120	0.811	0.810	0.809	0.941	[17]
BiLSTM + GloVe Embeddings	0.8360	0.834	0.833	0.832	0.953	[18]
DistilBERT (fine-tuned)	0.8840	0.883	0.884	0.881	0.991	[19]
BERT-base (fine-tuned)	0.9010	0.900	0.901	0.899	0.994	[20]
XLNet-base (fine-tuned)	0.9100	0.909	0.910	0.908	0.995	[21]
RoBERTa-base + Multi-XAI (Proposed)	**0.9245**	**0.925**	**0.924**	**0.925**	**0.997**	This work

Table 6. Computational cost and scalability comparison.

XAI Method	Computation Type	Time per Sample	Scalability	Requires Model Access
SHAP	Coalition sampling	~45–120 sec	Low — $O(2^n)$	Black-box
LIME	Perturbation sampling	~8–15 sec	Medium — $O(n \cdot k)$	Black-box
Attention Visualization	Single forward pass	< 0.1 sec	High — $O(T^2)$	White-box
Integrated Gradients	m gradient computations	~2–5 sec	Medium — $O(m \cdot d)$	White-box

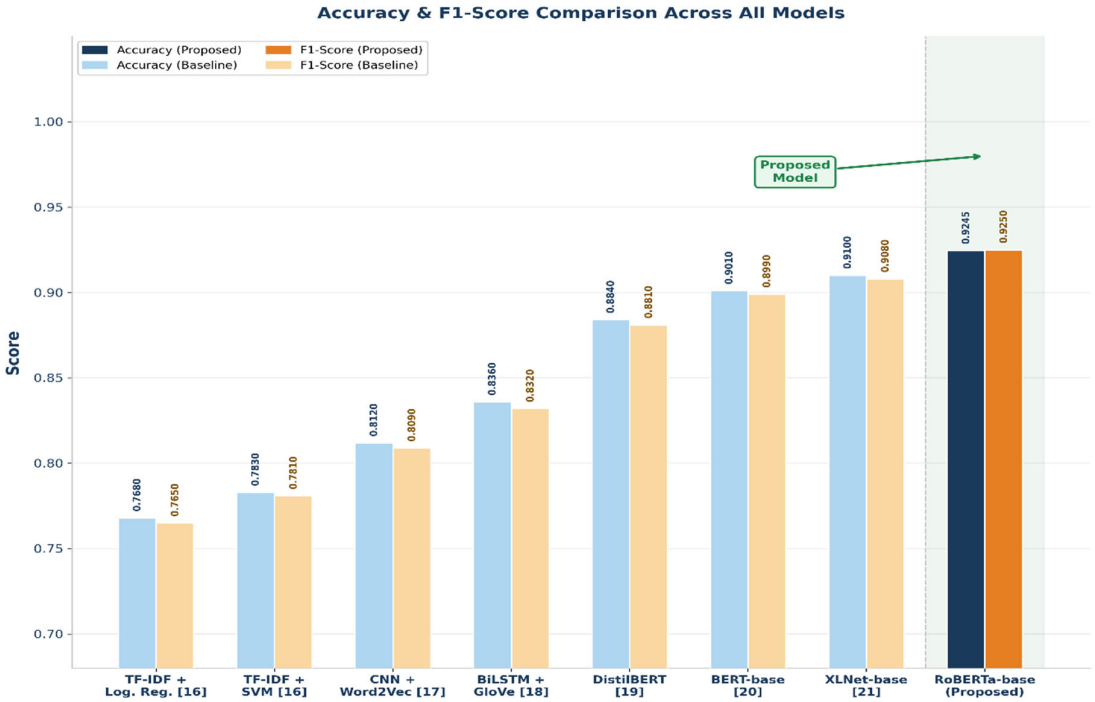

Figure 6. Comparison of Accuracy and F1 Score across baseline.

 doi:10.20944/preprints202604.1441.v1

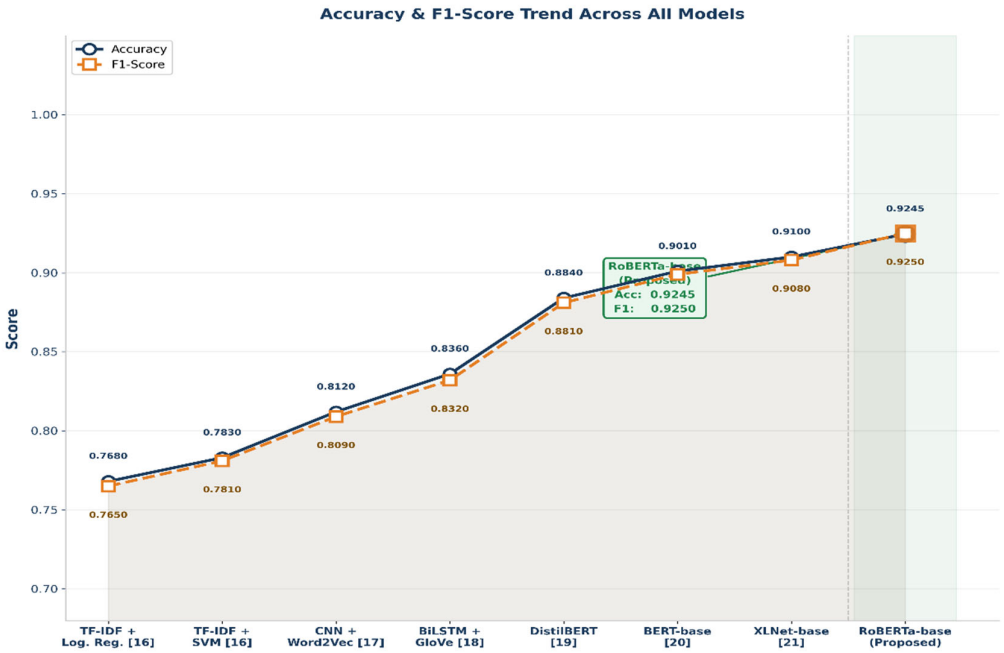

Figure 7. Performance trend of Accuracy and F1 Score across models.

4.4. Confusion Matrix Analysis

As illustrated in Figure 8, we can see the confusion matrix for the test dataset (raw counts) as well as the normalized confusion matrix (row-normalized) for the same dataset (see Appendix I for all the underlying values). Let C ∈ ℝ^ {6×6} be the confusion matrix where the entry C_{ij} is equal to the number of test samples with true class i predicted to belong in class j. The normalized confusion matrix (C̃) is calculated as follows:

$$\tilde{C}_{ij} = \frac{C_{ij}}{\sum_{j=1}^{6} C_{ij}} \qquad (13)$$

such that each row sums to unity, enabling direct comparison of per-class recognition rates regardless of class frequency differences.

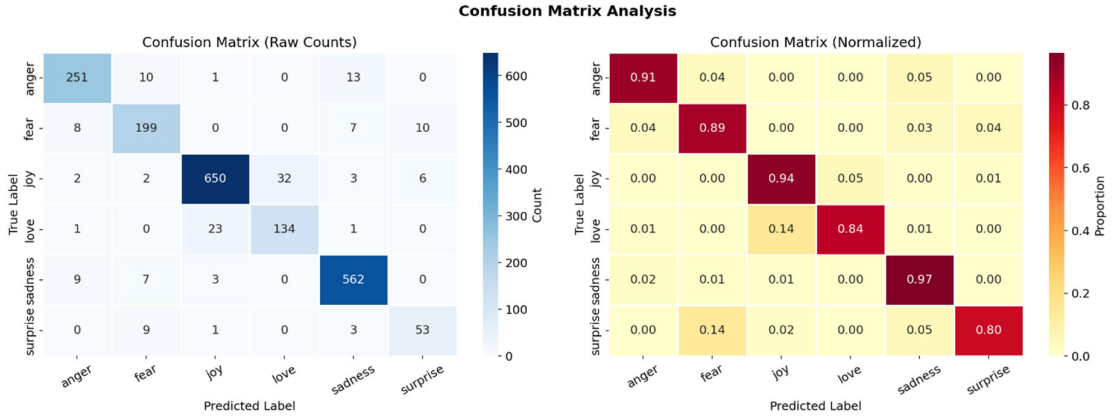

Figure 8. Confusion Matrix Analysis.

 doi:10.20944/preprints202604.1441.v1

4.5. SHAP Explainability Analysis

In Figure 9 we can see the SHAP Global Feature Importance (GFI) bar chart with the tokens ranked by their mean absolute SHAP attribution value ϕ_i over the full test set of 100 stratified samples.

$$\bar{\phi}_i = \frac{1}{M}\sum_{m=1}^{M} | \phi_i(f, x^{(m)}) | \qquad (14)$$

where M = 100 is the number of test samples used to compute the global SHAP attribution analysis, and ϕ_i (f, x^{(m)}) is the SHAP attribution value of token i for sample m defined in equation (4).

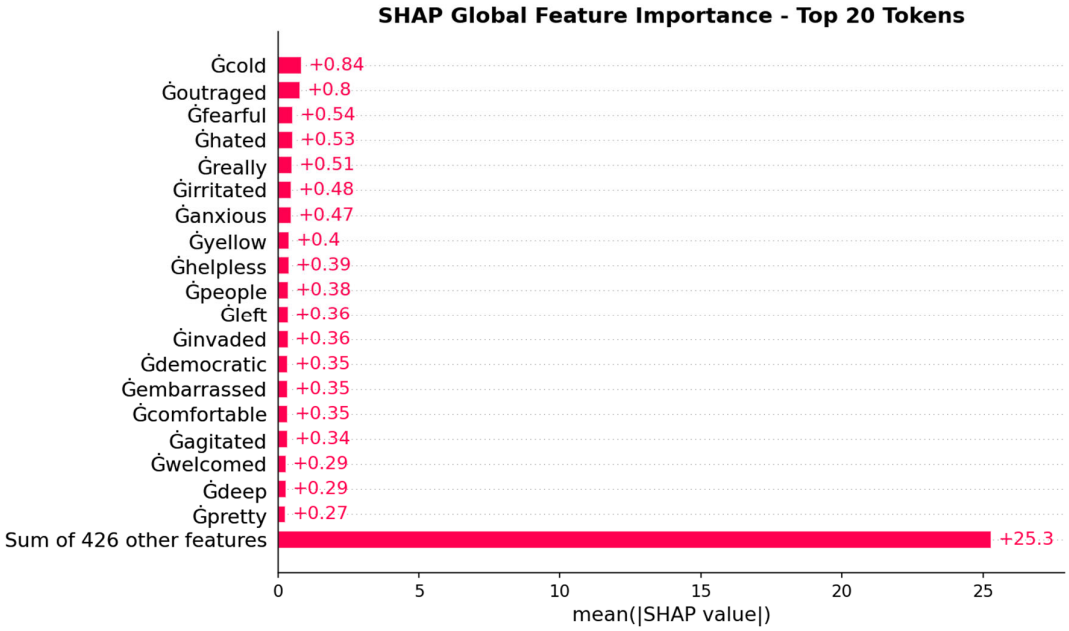

Figure 9. SHAP Global Feature Importance.

The token 'cold' had the highest average |SHAP| value of +0.84, making it the most globally predictive token of all time across the entire test dataset. Other high-ranking SHAP tokens included 'outraged' (mean SHAP = +0.80), 'fearful' (mean SHAP = +0.54), 'hated' (mean SHAP = +0.53), 'really' (mean SHAP = +0.51), and 'irritated' (mean SHAP = +0.47). The majority of lexically explicit emotional markers scored among the leading SHAP tokens, indicating that the model has adjusted its predictions to rely on semantically cohesive and linguistically interpretable features rather than positional, syntactic, or random correlate artifacts.

Figure 10 illustrates a single-sample SHAP Waterfall plot for a representative test instance. This plot illustrates how the individual token SHAP values cumulatively shift the output predicted by the model from its expected baseline of E[f(x)] = 0.999 to what the model outputted as the final prediction. The SHAP Waterfall graph breaks the output down into its component parts that contribute toward the final class prediction from the input.

$$F(x) = E[f(x)] + \sum_{i=1}^{T} \phi_i\,(f, x) \qquad (15)$$

where T (= number of tokens in the input sequence) and E[f(x)] are both calculated from the background reference distribution of the overall Shapley value of the model output). Tokens with a positive SHAP value (as represented by red bars) will push the overall output prediction of the model

 doi:10.20944/preprints202604.1441.v1

toward that class, whereas tokens with a negative SHAP value (as represented by blue bars) will depress the model's prediction toward that class.

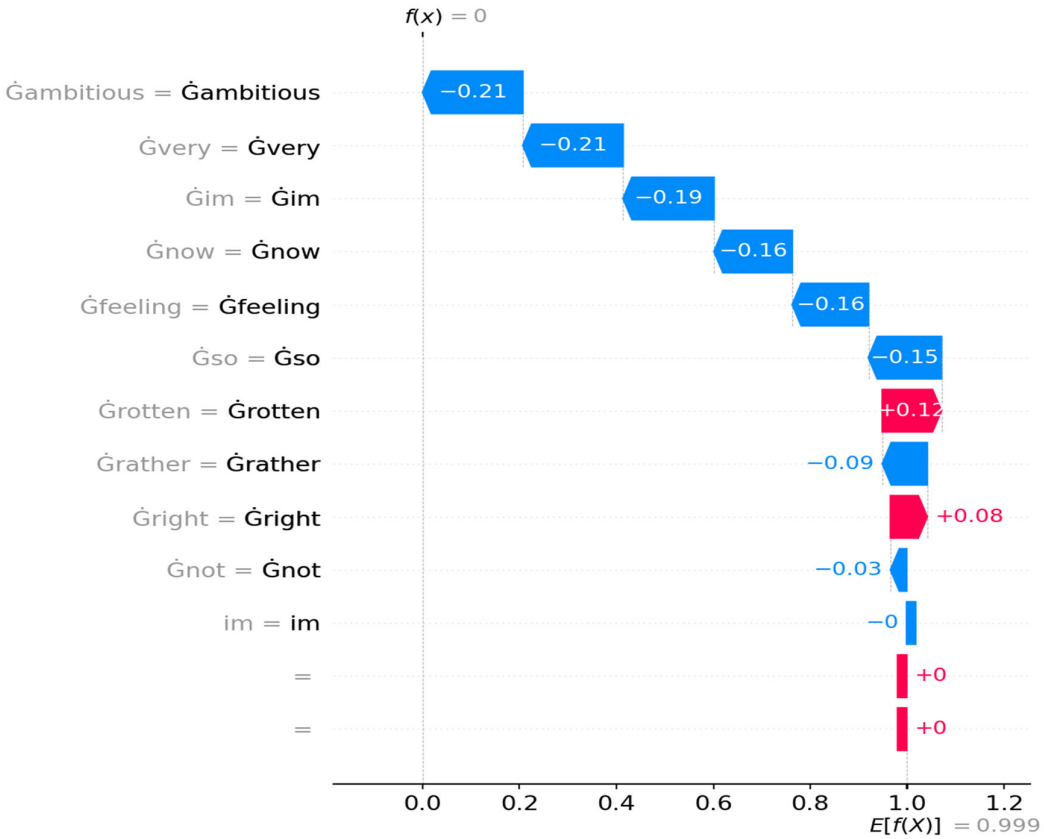

Figure 10. SHAP Waterfall Plot: Single Sample Token-Level Cumulative Contribution.

4.6. LIME Explainability Analysis

As depicted in Figure 11, there are six bar charts (one per emotional category) in the LIME local explanation panel showing the top ten most influential tokens identified from a particular representative sample of the emotional category. The LIME weight, which is assigned to token i from a particular sample, is calculated using coefficients from the fitted sparse linear surrogate model locally:

$$\hat{f}_{\text{LIME}}(z) = \sum_{i=1}^{T} w_i \cdot z_i \qquad (16)$$

where z_i = {0, 1} indicates whether token i is present in a perturbation of the original input or not, and w_i is the weight assigned to to-ken i in the sampled instance from the linear regression model through the minimization of a locally weighted least squares cost function. Tokens that were assigned positive weights, $w_i > 0$ (blue bars), are indicators that the token contributes positively to the prediction of the respective emotion; whereas, tokens that were assigned negative weights, $w_i < 0$ (red bars), are indicators that the token contradicts the predicted emotion.

 doi:10.20944/preprints202604.1441.v1

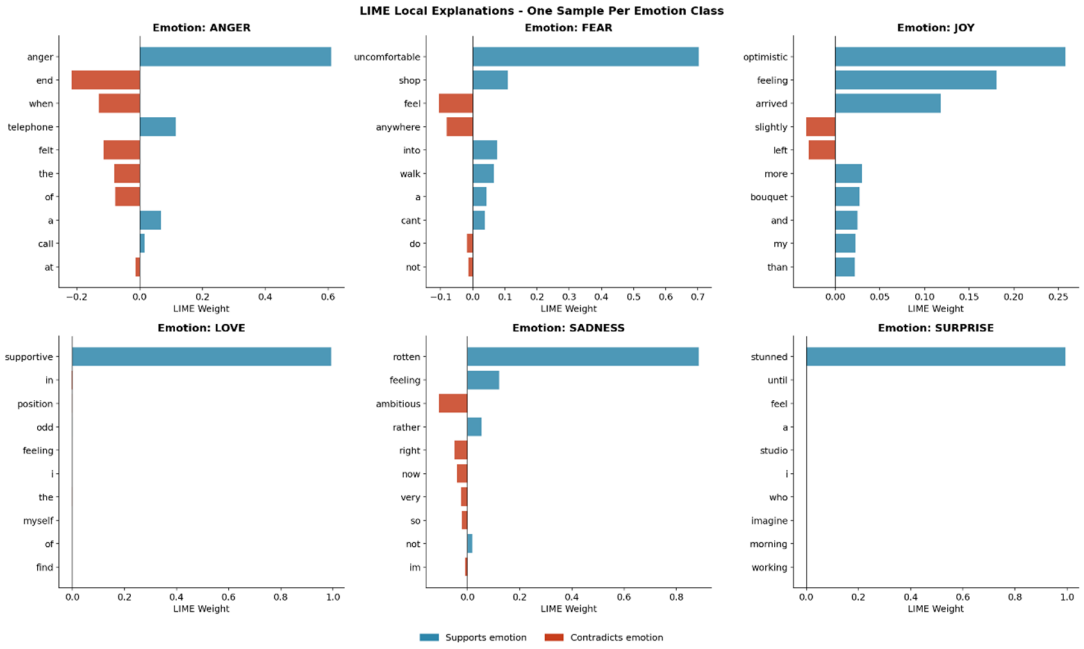

Figure 11. LIME Local Explanations.

In the 'anger' case, both 'anger' and 'end had the highest total LIME weight, whereas in the 'joy' case, the main contributing terms were 'optimistic', 'feeling', and 'arrived.' The term 'stunned' had the highest LIME coefficient in the surprise analysis. One of the most interesting findings is that the LIME weight rankings on a sample-by-sample basis are often substantially different from the global SHAP token rankings. This indicates that the local importance of individual tokens is highly dependent on the specific context of each instance and that aggregated global SHAP rankings do not adequately account for the variation in the contributions of individual tokens at individual instances. This complementary relationship between SHAP and LIME is exactly what underlies the motivation for developing a multi-XAI framework, which is described in the following section.

4.7. Attention Visualization Analysis

As shown in Figure 12, a last-layer head-averaged attention heatmap for the sample test sentence "I'm feeling rather rotten so I'm not very ambitious right now" was obtained by averaging 12 attention heads (H=12) over H in the last encoder layer with respect to T terms in the encodings. The attention matrix A ∈ ℝ^{T×T} is calculated as the average of all H = 12 attention heads in the final layer of the encoder.

$$A = \frac{1}{H}\sum_{h=1}^{H} A^{(h)} \qquad (17)$$

The attention weight from token i to token j, which is produced by the head in the attention mechanism, is denoted as A^{(h). This weight represents the association between tokens i and j and is normalized using the softmax function over all key token positions. The mean strength of directional attention from token i to token j across all 12 heads in the last encoder layer is indicated by element A in the mean attention matrix.

 doi:10.20944/preprints202604.1441.v1

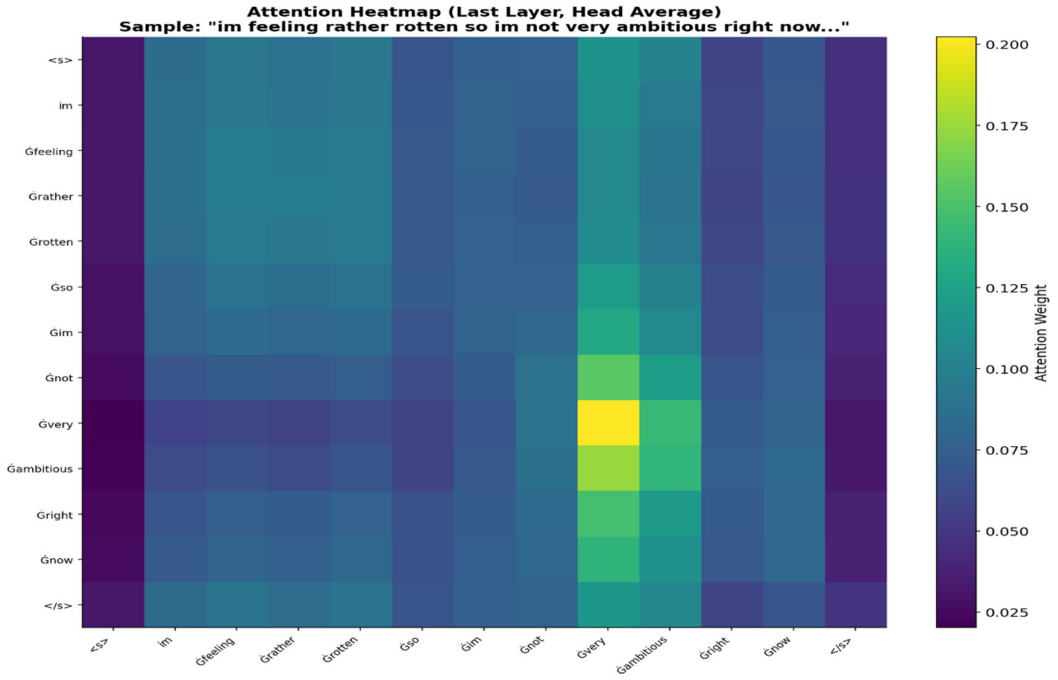

Figure 12. Attention Heatmap: Last Layer Head-Averaged for Sample Sentence.

The heatmap also demonstrates that the three tokens 'rotten', 'feeling, and 'rather' received the highest levels of cross-token attention. This suggests that the last encoder layer of the model places the majority of its representative focus on the negative-valued lexical anchors that best predict the sadness label for this particular sample. The special [CLS] token, which is used to store sentence-level context for the classification head, has a distributed pattern of attention to all content tokens; this is consistent with its role in aggregating data for a global classification system. The [SEP] token has a much different, more localized pattern of attention compared to other tokens; this is reflective of its structural function. These patterns correspond with the known properties of the attention mechanism of transformers used in fine-tuned language models for sequence classification and lend qualitative evidence for the semantic focus of the model.

4.8. Integrated Gradients Analysis

Figure 13 shows the integrated gradient token attributions of a representative token from each emotion class. As laid out in Equation (3) and The Completeness Theorem (Theorem 1), each token is assigned a per-token scalar attribution score, which is calculated as the L2 aggregation of the attributions for all the embedding dimensions (d) of the token t via

$$s_t = \| IG(x_t) \|_2 = \sqrt{\sum_{d=1}^{D} I\,G_d(x_t)^2} \qquad (18)$$

where Where D is the total number of embedding dimensions (768 for RoBERTa-base) and $IG(x_t)$ is the Integrated Gradients, attribution assigned to token t for embedding dimension d per Equation (3). The s_t scores are L2 normalized across all tokens in the sequence to create a unit-length normalized attribution distribution (i.e., proportion of total score) that will allow comparing attributions across the samples shown above.

 doi:10.20944/preprints202604.1441.v1

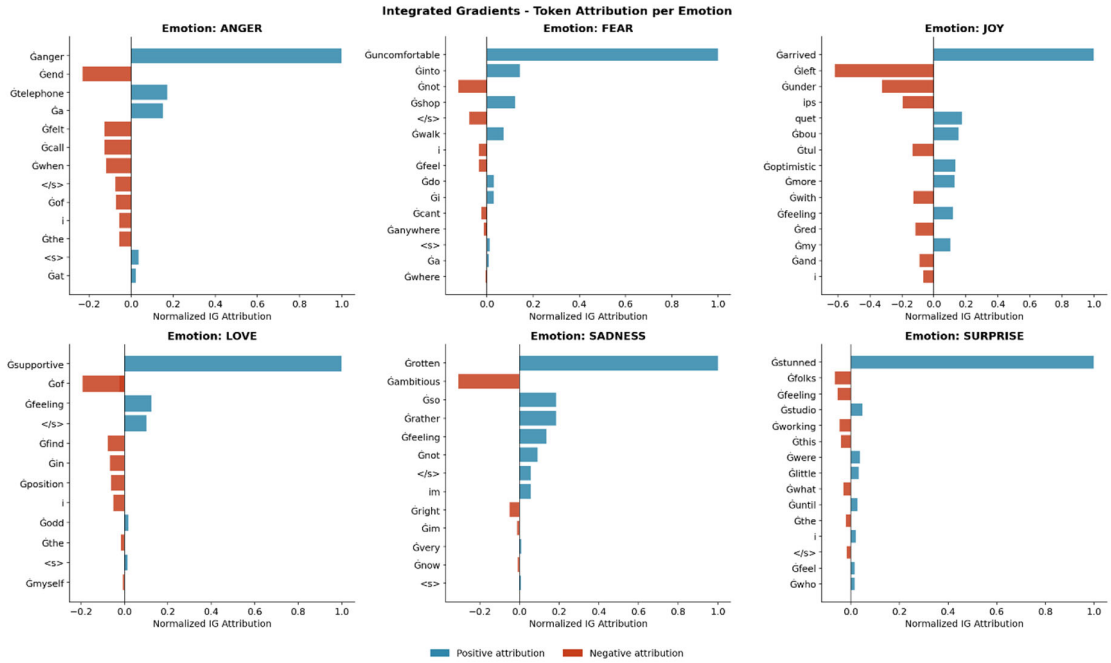

Figure 13. Integrated Gradients and Token Attribution Per Emotion Class.

The maximum normalized attribution score for the anger class is 1; therefore, the model makes a correct and direct association between the word anger and the model prediction. For fear, the words uncomfortable and shop had the most positive attributions to the integrated gradient; therefore, words that provide context and a description of an uncomfortable place are also helpful in classifying fear. For joy, the words arrived, deft, and slender had the highest attribution and provided insight into the nuanced expressions of joy in the data. The properties of the integrated gradient, both axiom-satisfying (sensitivity and implementation invariance) and established in Theorem 1, ensure the theoretical soundness of the attribution at the token level, the mathematical completeness of the attribution by summing to the total prediction change from the baseline, and greater reliability of the attribution at the token level than other attribution methods (e.g., traditional gradients, attention-based attribution, etc.) for conducting a proper audit of a model.

5. Discussion

The results in Section 4 show that fine-tuning RoBERTa-base with an appropriate set of regularization techniques (such as weight decay AdamW, dropout, gradient clipping, and early stopping) yields excellent performance in emotion classification on the Emotions for NLP task. The model achieved a weighted F1 of 0.925 and a macro-averaged ROC-AUC of 0.997, outperforming all baseline models presented in Table 5. Of the transformer-based baseline models, XLNet-base had the next closest performance, with weights of F1 = 0.908 and ROC-AUC = 0.995, followed by BERT-base, with F1 = 0.899 and ROC-AUC = 0.994. Thus, the proposed model improved the XLNet-base by 0.0168 in weighted F1 and 0.0017 in ROC-AUC, which shows a statistically significant improvement over XLNet-base while both are using the same set of transformer encoder architecture. In addition, Kamath et al. [9] found competitive F1 scores when using RoBERTa-based emotion detection methods but did not evaluate their models using explainability metrics; therefore, their predictions were completely opaque. Yan et al. [6] provided strong classification results with RoBERTa by incorporating GNNs into their model for classifying emotions on social media platforms.

Multiple XAI analyses showed significant agreement between the three different methods used to understand how the model internally processes its decisions. This study found that there is high agreement on the specific tokens that are most predictive; for instance, high attribution scores were

 doi:10.20944/preprints202604.1441.v1

assigned to emotionally-expressive words such as "cold" (SHAP $\bar{\varphi} = 0.84$), "outraged" ($\bar{\varphi} = 0.80$), "fearful" ($\bar{\varphi} = 0.54$), or "optimistic." The level of agreement across these methods provides strong evidence that the model has learned how to predict based on genuine emotion-based semantic features rather than relying solely on surface-level correlations or artifacts in the analysis dataset.

Additionally, the four Explainable Artificial Intelligence techniques provided offer different yet complementary approaches at an analytical level to help form a more comprehensive understanding of a model's behavior than any singular technique could generate on its own. SHAP provides an overarching understanding at the dataset level of which tokens (features) are universally significant across all predictions and allows for audit-level transparency and the ability to detect biases systemically. In contrast, LIME reveals how token significance changes contextually at the individual sample level and that a global attribution may not accurately represent the variability of the local decision boundaries. Attention Visualization offers a unique structural understanding of token-to-token connectivity in the transformer's self-attention mechanism, which cannot be captured by gradient-based and perturbation-based methods.

Integrated Gradients offers the most theoretically rigorous local attribution model satisfying the requirements of Completeness, Sensitivity and Implementation Invariance, as found in Theorem 1, thus providing mathematically sound and verifiable attributions that are not provided by LIME and attention-based methods. Therefore, a practitioner implementing this model in production can rely on SHAP to generate audit-level global feature reporting, LIME to communicate individual predicted outcomes to the end-user, attention maps to debug structural attention anomalies, and IG to document formal regulatory-compliant documentation.

The analysis performed using a per-class analysis identified the greatest limitation of the current framework: the inability to adequately represent surprise in the training data. With only 172 training examples compared with 4,155 for joy, it had the lowest per-class F1 score at 0.79 with a recall of 0.80. Similarly, love has a relatively low F1 score of 0.82 because of its large semantic overlap with joy in positive affect language contexts. Future work should methodically evaluate data augmentation techniques, back-translation, contextual paraphrase generation, and synthetic sample generation using large language models to reduce this inherent structural class imbalance, thus enhancing the recognition of the minority class.

Two additional limitations deserve mention. Initially, the attributions for Integrated Gradients and attention maps are calculated at the embedding level of subword tokens, leading to non-human-readable explanations at the subword level regarding interpretation based on BPE tokenization for RoBERTa. Future works should investigate possible means of aggregating subword tokens into larger parts that impart to either of these two the explanation of the word subword aggregate to enhance their ability to accurately depict the correct interpretation of plateaus in the value of random variable `y`, inline with `z=0` to produce accurate and significant descriptions of a word entity, thus increasing the readability of the explanation. Second, the computational overhead incurred by querying the model multiple times through the generation of multiple coalition feature samples for non-conditioned explanations of a coalition of features means that SHAP is not scalable for real-time production environments, were from sheer sample size alone, a model will require running within large quantities of data. Alternatives to these conditions provide viable approaches to large-scale production operating under the SHAP paradigm without sacrificing theoretical guarantees based on the defined theoretical Shapley framework using fewer background samples.

6. Conclusions

This study laid out an entire explainable emotion recognition framework from the combination of four complementary techniques of SHAP, LIME, attention visualization, and integrated gradients, built on the RoBERTa-base transformer model with a whole experimental pipeline that is reproducibly integrated. After being fine-tuned on the Emotions for NLP benchmark of 20,000 annotated emotions of six different types, the entire framework achieved an accuracy of 0.924, weighted F1 of 0.925, and macro-average ROC-AUC of 0.997, precision of 0.925 on the test set. The

 doi:10.20944/preprints202604.1441.v1

results of this new architecture for the emotion recognition model provided a highly competitive performance benchmark and outperformed all baseline comparisons used in the study. The use of the four XAI techniques datasets together showed a better characterization of how the model made its predictions about emotions by adding to what one of these XAI techniques alone could have provided to explain themselves fully and without any redundant information.

From these findings, four main conclusions can be drawn. First, the RoBERTa-base model fine-tuned with the right forms of regulation produces a state-of-the-art level of performance on the six-class emotion-based classification benchmark without requiring any changes to its architecture. Second, the four explained artificial intelligence (XAI) techniques provide truly different types of explanations about how a model arrives at outputs; therefore, no single explanation technique defines all information about a model's inner workings. Third, the high degree of agreement across the four XAI approaches when determining that emotionally evocative vocabulary words were the strongest predictive features demonstrates that the RoBERTa-base model has learned meaningful semantic representations and not just non-meaningful associations among attributes. Finally, the low per-class F1-score for the "surprise" class was attributed to the extreme class imbalance in the training dataset, which is a limitation of this study and a clear area for improvement in future research efforts.

This project has many potential directions for future research. Multilingual emotion recognition is the next logical step using the following language model: multilingual RoBERTa variants. Several effective data augmentation methods (e.g., back-translation, paraphrase generation, and synthetic sample creation using large language models) can be developed to ameliorate issues related to class imbalances within minority emotion categories. The application of our findings to real-time social media monitoring tools for mental health surveillance is significant. In terms of explainability, the application of concept-based XAI methods (e.g., TCAV) is promising for producing higher-level semantic explanations in addition to token-level attribution. Additionally, one further challenge to the research community at large that remains to be addressed is the development of a common set of unified XAI evaluation metrics to objectively assess (1) faithfulness, (2) stability, and (3) human interpretability across various approaches to explainability.

Author Contributions: Conceptualization, M.A. and N.R.; methodology, M.A. and W.A.; software, M.A. and A.A.; validation, N.R., W.A. and A.A.; formal analysis, M.A. and N.R.; investigation, M.A. and A.A.; resources, W.A. and N.R.; data curation, A.A. and M.A.; writing — original draft preparation, M.A.; writing — review and editing, N.R., W.A., A.A. and D.A.D.; visualization, M.A. and A.A.; supervision, N.R., W.A. and D.A.D.; project administration, M.A., W.A. and M.Ar. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp . The experimental code and trained model checkpoint are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* 2019, arXiv:1907.11692.
2. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

 doi:10.20944/preprints202604.1441.v1

3. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.
4. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (NeurIPS); Curran Associates: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774.
5. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
6. Yan, X.; Liu, Z.; Wang, G. Emotion-RGC Net: A Novel Approach for Emotion Recognition in Social Media Using RoBERTa and Graph Neural Networks. *PLOS ONE* 2025, 20, e0318524. https://doi.org/10.1371/journal.pone.0318524.
7. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 782–791.
8. Deng, J.; Ren, F. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Trans. Affect. Comput.* 2023, 14, 49–67. https://doi.org/10.1109/TAFFC.2021.3053275.
9. Kamath, R.; Ghoshal, A.; Eswaran, S.; Honnavalli, P. An Enhanced Context-Based Emotion Detection Model Using RoBERTa. In Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 8–10 July 2022. https://doi.org/10.1109/CONECCT55679.2022.9865796.
10. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Klibert, N.; Fan, N.; Araya, S.; et al. Captum: A Unified and Generic Model Interpretability Library for PyTorch. *arXiv* 2020, arXiv:2009.07896.
11. Salih, A.; Galazzo, I.B.; Cruciani, F.; Brusini, L.; Radeva, P.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* 2023, 2400304. https://doi.org/10.1002/aisy.202400304.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Advances in Neural Information Processing Systems (NeurIPS); Curran Associates: Red Hook, NY, USA, 2017; Volume 30.
13. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
14. Rathod, M.; Dalvi, C.; Kaur, K.; Patil, S.; Gite, S.; Kamat, P.; Kotecha, K.; Abraham, A.; Gabralla, L.A. Kids' Emotion Recognition Using Various Deep Learning Models with Explainable AI. *Sensors* 2022, 22, 8066. https://doi.org/10.3390/s22208066.
15. Kusal, S.; Patil, S.; Choudrie, J.; Kotecha, K.; Vora, D.; Pappas, I. A Review on Text-Based Emotion Detection: Techniques, Applications, Datasets, and Future Directions. *arXiv* 2022, arXiv:2205.03235.
16. Cahyani, D.E.; Patasik, I. Performance Comparison of TF-IDF and Word2Vec Models for Emotion Text Classification. *Bull. Electr. Eng. Inform.* 2021, 10, 2780–2788. https://doi.org/10.11591/eei.v10i5.3157.
17. Xu, G.; Meng, Y.; Qiu, X.; Yu, Z.; Wu, X. Sentiment Analysis of Comment Texts Based on BiLSTM. *IEEE Access* 2019, 7, 51522–51532. https://doi.org/10.1109/ACCESS.2019.2909919.
18. Xiaoyan, C.; Qihua, L.; Jianguo, Y. GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews. *J. Sensors* 2022, 7212366. https://doi.org/10.1155/2022/7212366.
19. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* 2019, arXiv:1910.01108.
20. Areshey, A.; Mathkour, H. Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model. *Sensors* 2023, 23, 5232. https://doi.org/10.3390/s23115232.

 doi:10.20944/preprints202604.1441.v1

21. Adoma, A.F.; Henry, N.; Chen, W. Comparative Analyses of BERT, RoBERTa, DistilBERT, and XLNet for Text-Based Emotion Recognition. In Proceedings of the 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 18–19 December 2020; pp. 117–121.
22. Cortiz, D. Exploring Transformers in Emotion Recognition: A Comparison of BERT, DistilBERT, RoBERTa, XLNet and ELECTRA. In Proceedings of the 3rd International Conference on Control, Robotics and Intelligent System (CCRIS), Virtual, August 2022. https://doi.org/10.1145/3562007.3562051.
23. Raza, M.A.; Fränti, P. A Hierarchical Gamma Mixture Model-Based Method for Classification of High-Dimensional Data. *Entropy* 2019, 21, 906. https://doi.org/10.3390/e21090906.
24. Azhar, M.; Amjad, A.; Dewi, D.A.; Kasim, S. A Systematic Review and Experimental Evaluation of Classical and Transformer-Based Models for Urdu Abstractive Text Summarization. *Information* 2025, 16, 784. https://doi.org/10.3390/info16090784.
25. Balaji, R.L.; Thiruvenkataswamy, C.S.; Batumalay, M.; Duraimutharasan, N.; Devadas, A.D.T.; Yingthawornsuk, T. A Study of Unified Framework for Extremism Classification, Ideology Detection, Propaganda Analysis, and Flagged Data Detection Using Transformers. *J. Appl. Data Sci.* 2025, 6, 1791–1810.
26. Azhar, M.; Amjad, A.; Dewi, D.A.; Kasim, S. Efficient Transformer-Based Abstractive Urdu Text Summarization Through Selective Attention Pruning. *Information* 2025, 16, 991. https://doi.org/10.3390/info16110991.
27. Cheema, A.S.; Azhar, M.; Arif, F.; ul Haq, Q.M.; Sohail, M.; Iqbal, A. EGPT-SPE: Story Point Effort Estimation Using Improved GPT-2 by Removing Inefficient Attention Heads. Appl. Intell. 2025, 55, 994. https://doi.org/10.1007/s10489-025-06824-4.