International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1014-1037
https://doi.org/10.70917/ijcisim-2026-3483
Article

Enhanced Hybrid Emotion Recognition System: Addressing Class Imbalance and Emotion Hierarchies through Feature-Rich Deep Learning

Priya George A[1*], Roshan Fernandes[2]

[1]NMAM Institute of Technology, Nitte, Visveswaraya Technological University, Belagavi-590018, India
[1]School of Engineering, St Aloysius (Deemed to be University), Mangalore, India
[2] Research Supervisor, VTU Research Centre, NMAM Institute of Technology, Nitte, Visveswaraya Technological University, Belagavi-590018, India
Email: [1*] annpriya.2013@gmail.com, [2] roshan_nmamit@nitte.edu.in
*Corresponding Author

Abstract: — The task of detecting emotions at a detailed level in written text continues to be difficult because of three main issues which include the large difference between the number of examples in each category and insufficient features and intricate connections between different emotions. The current transformer models which analyze text data achieve unsatisfactory results by producing F1 scores which are less than 0.54 when evaluated on the GoEmotions benchmark. The research solves these problems by using a three-stage emotion recognition system which identifies 28 different emotions. The research introduces an EmotionAwareRebalancer system which uses transformer-based text hybridization to create realistic synthetic data that solves class imbalance problems by reducing the 81:1 ratio to 4:1. The research presents three main contributions which include a 172-dimensional feature extraction system that combines contextual embeddings with linguistic markers and lexicon-based signals and stylistic patterns and a deep learning system that uses multi-level attention mechanisms and hierarchical loss functions to model psychological emotion relationships. The research method uses DistilRoBERTa embeddings with cosine similarity thresholds (0.85) to preserve semantic meaning in synthetic data generation and trains the model with focal loss (γ=2) and hierarchy-constrained regularization. The experimental results show that the proposed method reaches 0.624 F1 score on test data which surpasses RoBERTa-large (0.503) by 24.1% and EmoBERT (0.523) by 19.3%. The results show that minority class performance reaches significant improvement because grief increases from 0.21 to 0.541 F1 and relief increases from 0.18 to 0.540 F1. The ablation study demonstrates that model rebalancing leads to a 0.066 F1 score improvement and the hierarchical loss function adds another 0.081 F1 score improvement. The framework shows how it can be used in real-world applications to monitor mental health and analyze customer feelings and develop human-computer interfaces which need advanced emotional processing capabilities.

Keywords: — Emotion Recognition, Class Imbalance, Multi-label Classification, Semantic Data Augmentation, Hierarchical Classification, Attention Mechanisms, Natural Language Processing, Affective Computing

1. INTRODUCTION

Emotion recognition from Text document is becoming increasingly important with developments in graphic design psychology, user experience assessment, and marketing. With the rapid growth of digital communication, the need to recognize and understand the emotions behind the messages has grown for many companies and organizations that want to figure out how to improve customer relations [30]. Most emotion analysis techniques use statistical and rule-based approaches which tend to be ineffective when dealing with human emotions and language [31]. To solve these problems, more and more researchers are applying machine learning (ML) and deep learning (DL) implementing algorithms that automatically classify emotions in text documents. The major hurdles that prevent accurate recognition of emotion still persist, even with the global machine learning [32] advancements such as contextual dependencies, linguistic variations, and emotional expression's subjectivity. Emotion recognition in Natural Language Processing

(NLP) has always been and continues to be a challenge. The nuanced essence of emotions alongside contextual emphasis makes it intricate for the algorithm to detect the sentiment of textual data [33].

A. Literature Review

The study on emotion detection from text data has progressed significantly in the integration of traditional and modern deep learning approaches [34]. In the past, sentiment analysis was based on lexicon and some basic statistical methods [35]. These approaches usually neglected the interrelations and other complexities associated with emotions expressed in text. Consequently, the use of machine learning algorithms like Support Vector Machines and Naive Bayes was adopted as they proved to be more scalable and adaptable to different environments. Yet, there remains stubborn ambiguity and context deep emotions which these primitive methods could not process [36].

Numerous studies have looked into the use of ML technologies for emotion analysis. Uma Rani et al. [1] examined a variety of supervised learning approaches, including Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Random Forests. Their analysis proved that deep learning models, particularly Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) substantially exceed traditional machine learning methods in precision, recall, and accuracy. Likewise, Priyadarshini and Cotton [2] developed a new hybrid model and optimized it with grid search for the LSTM-CNN with improved classification capabilities They stated that social media networks constitute a rich resource for sentiment detection particularly during important milestones like the COVID-19 pandemic. Hansun et al. [3] studied sentiment on Twitter for the PPKM policy of the Indonesian government. The results showed that most sentiments expressed were neutral at 58.07%, while positive and negative sentiments accounted for 27.12% and 14.81%, respectively. To further boost accuracy of classification, they used deep learning with an LSTM approach, which resulted in an accuracy of 92.59%. In multilingual emotion analysis, Chundi et al. [4] addressed the challenges posed by Kannada-English code-switching. Their proposed lexicon-based approach, NBLex, demonstrated improved efficiency in processing mixed-language texts, achieving 83.2% accuracy and an F1-score of 83%. to enhance the sentiment classification process, researchers have explored multi-task learning and feature abstraction techniques. Park and Kim [5] introduced a stepwise multi-task learning model incorporating RoBERTa and Bi-LSTM for fine-grained aspect-based sentiment analysis. Their model effectively extracted emotion -related aspects, improving generalization across datasets.

Z-Qi [6] focused on dynamic emotion analysis using ON-LSTM and attention mechanisms. The study found that integrating label semantics improved classification accuracy and F1-score, demonstrating that ON-LSTM-based models effectively capture hierarchical text structures for emotion recognition. Recent advancements emphasize multimodal emotion analysis, integrating text and visual data for enhanced performance [38]. Habib et al. [7] examined deep learning fusion techniques, combining RoBERTa with EfficientNet-b3 and ResNet50. Their multimodal approach achieved an accuracy of 75%, outperforming single-modality models. Khalid et al. [8] tackled emotion analysis of deepfake-related tweets, introducing a sentiment majority voting classifier (SMVC). Their hybrid approach employed lexicon-based models (TextBlob, VADER, AFINN) and transfer learning with LSTM and Decision Trees. The logistic regression model achieved a remarkable 98.9% accuracy, proving the effectiveness of hybrid feature engineering. Obiedat et al. [9] addressed imbalanced sentiment data in customer reviews by integrating SVM with Particle Swarm Optimization (PSO) and various oversampling techniques. Their findings showed that the PSO-SVM combination improved classification metrics, demonstrating robustness against data imbalance [40].

Wang [10] analyzed different techniques for word embedding and their effectiveness at emotion recognition using sentiment classification models on the IMDB dataset. The research focused on Rand, Static, Non-static and multichannel Word Embeddings, measuring their relevance to the functioning of the neural networks involved. The findings showed that the randomly initialized model resulted in the best performance, as it was able to classify emotions strongly even without pre-trained embeddings. Shen [11] proposed an interaction deep learning model called BERT-LLSTM-DL for sentiment analysis of recent Chinese literature works. This model uses BERT for language representation and LSTM for sequential learning and achieves high accuracy on sentiment classification tasks. The study noted contextual understanding is key to achieving the desired results in sentiment analysis. Yan and Wang [12] designed the Attention-BiLSTM model for sentiment classification of micro-blog data. This model incorporated the CBOW method of word embedding in Word2Vec and an attention mechanism that re-evaluates the weight assigned to each input vector. The proposed model outperformed other deep learning models like Text CNN and BiLSTM in classification accuracy and generalization ability. Mutinda et al. [13] leveraged sentiment lexicons with pre-trained BERT embeddings and developed a sentiment analysis model based on LeBERT and CNN to enhance sentiment classification of text reviews. In this hybrid model, a sentiment lexicon was integrated in the framework of BERT to improve classification.

LeBERT demonstrated superior performance over the existing models in binary sentiment classification tasks with an impressive F-measure score of 88.73%. This value is very close to the optimal score considered achievable. In their paper, Dashitpour et al. [14] did sentiment analysis in Persian languages using CNN and LSTM models. They noted how deep learning methods were better or worse than their classical statistical machine learning counterparts and concluded that LSTM was the best supported method for classifying movie reviews in Persian. Lakshmi [15] presented a review of the progress made in a particular area of deep learning – that of sentiment analysis – and tackled concerns about model interpretability, domain of application, and ethics. This study focused on developing cross-domain and cross-lingual sentiment analysis models to address the limitations of deep learning techniques.

Wang [16] analyzed the performance of different deep learning models in sentiment analysis tasks. The study compared CNNs, RNNs, and BERT-based models, concluding that BERT achieved the highest accuracy, while CNNs performed better than RNNs in handling sentiment classification. Manikandan et al. [17] proposed a deep sentiment learning approach for similarity recommendations on Twitter data. The study used CNN and RNN to analyze multimedia content and textual data. The results indicated that deep belief networks performed effectively in classifying Twitter sentiment data. Mandhasiya et al. [18] explored hybrid deep learning models incorporating CNN, LSTM, and BERT for sentiment analysis of Indonesian political discussions on YouTube. The study demonstrated that hybrid models outperformed standalone deep learning models in sentiment classification [70][71].

Despite significant advancements in deep learning for emotion recognition, several challenges persist, limiting its effectiveness across diverse applications [19]. Traditional word embedding techniques, such as Word2Vec and GloVe, struggle to capture contextual dependencies, affecting classification accuracy in complex linguistic structures [41]. While transformer-based models like BERT and hybrid architectures (e.g., BERT-CNN, Attention-BiLSTM) have improved emotion prediction, they require extensive computational resources and large annotated datasets, restricting their deployment in resource-limited environments [20]. Issues like semantic loss, model interpretability, and domain adaptation further complicate sentiment classification, especially in nuanced textual contexts like literature and microblogging platforms. Thus, there is a need for optimized, interpretable, and resource-efficient emotion recognition frameworks that enhance accuracy while addressing cross-domain adaptability and computational constraints [21].

B. Problem Statement

Emotion recognition is a non-trivial task where three factors complicate the issue, namely the intricate way emotions are articulated, the relationship between emotions and their meaning, and the very large class ratio in the available datasets [22]. Most of the current methods are based on rudimentary feature extraction methods which do not consider the contextual and linguistic patterns, emotionality markers, or a stylistic marker, or any other aspect metaphoric to the taxonomy of skeletal language. Besides, existing deep learning systems for emotion classification usually consider emotions as isolated concepts, which breaks down psychologically defined hierarchies and relationships of emotion and results in incorrect classification. The GoEmotions dataset clearly embodies these issues with its 28 disparate emotion categories and extreme imbalance rate of 81:1, since most models cannot generalize to the minority emotion classes, yet perform strongly across the emotion spectrum. This study aims to solve these problems by developing an advanced hybrid emotion recognition system which has novel feature extraction methods and an emotion-aware deep learning architecture which is meant to model the hierarchical structure of emotions, mitigate class imbalance, and utilize rich linguistic features of emotions in the text.

C. Main Contributions

This study makes several significant contributions to the field of emotion recognition in text.

- First, we introduce a novel EmotionAwareRebalancer that addresses the extreme class imbalance challenge (81:1 ratio) through semantic-guided hybridization, effectively reducing it to a manageable 4:1 ratio while preserving the linguistic characteristics of minority emotion classes.
- Second, we develop a comprehensive feature extraction framework that captures the multidimensional nature of emotional expression by integrating contextual embeddings with emotion-specific linguistic markers, lexicon-based features, and structural-stylistic patterns, resulting in a rich 172-dimensional representation.
- Third, we propose an innovative emotion-aware deep learning architecture with specialized components including an emotion attention block, feature type gating mechanism, emotion hierarchy module, and

emotional aspect model, all designed to model the complex psychological relationships between different emotional states.

- Fourth, we implement a hierarchical loss function that enforces emotion relationship constraints during training, improving the model's ability to distinguish between closely related emotions.
- Our approach achieves state-of-the-art performance with validation F1 scores of 0.75 and test F1 scores of 0.624 across all 28 emotion categories in the GoEmotions dataset, outperforming previous benchmarks while maintaining interpretability through its hierarchical emotion modeling approach.

2. PROPOSED METHODOLOGY

The proposed methodology introduces a comprehensive three-phase approach to emotion recognition in text as shown in figure 1. Text input data is collected from GoEmotions dataset and performed exploratory data analysis (EDA) to understand the insights of the dataset. The Phase 1 focuses on preprocessing and class imbalance mitigation, where we implement a novel EmotionAwareRebalancer that addresses the extreme 81:1 imbalance ratio inherent in emotion datasets. This component employs semantic-guided hybridization techniques that intelligently combine text samples based on emotional similarity and hierarchical relationships, creating synthetic examples that preserve the linguistic characteristics of rare emotion classes. Through this approach, we effectively reduce the imbalance ratio to approximately 4:1, creating a more balanced training distribution without sacrificing the semantic integrity of the emotional content.

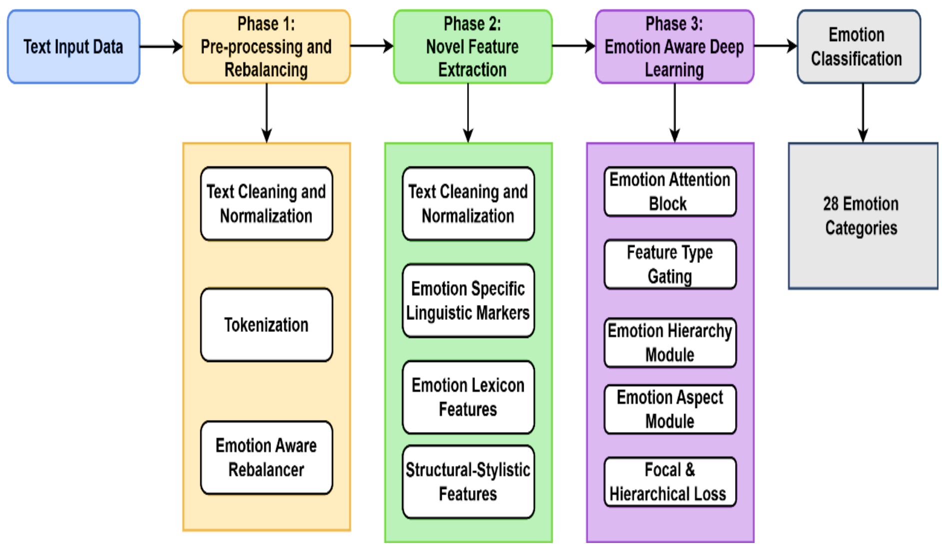

Figure 1: Proposed Methodology

In phase 2, we detail the novel feature extraction system designed especially for text with emotional expression and discuss its multidimensional nature. Unlike traditional methods which only utilize context embeddings, our method captures and combines four feature types. First, we have context embeddings, which capture the meaning of a text. Second is the emotion-specific linguistic features consisting of syntactic patterns related to emotional expression. Third, we have features identified from psychology's emotion taxonomic structures called lexicon features. Lastly, we have structural-stylistic features accompanied by different emotional patterns of writing. All of these features are integrated into a complete set of 172 dimensions which enable the deep learning model to capture rich emotional signals, not available through standard embedding methods [24]. In phase 3, we employ our emotion-aware deep learning architecture developed to capture the intricate features of emotions. The architecture integrates several new elements: An emotion attention block that concentrates on emotion relevant portions of text representations, a feature type gating block that assigns importance to certain types of features depending on their relevance to specific emotions, an emotion hierarchy module that models the psychological interrelationships among emotions, and an emotional aspect model that embodies the diverse forms of expression of emotion. These blocks are trained using a combination of focal loss to counter the remaining class imbalance, and hierarchical loss to impose emotion relation boundaries [43].

Data Collection

Our research employs the GoEmotions dataset, a revolutionary tool containing 58,000 annotated Reddit comments corresponding to 27 different emotions plus a neutral category [69]. This set is a monumental improvement in comparison to earlier emotion classification resources which were mainly limited to the six basic emotions introduced in 1992, or to small scale domain specific samples. GoEmotions' taxonomy was designed based on psychological principles, but in a rather practical way, containing 12 positive, 11 negative, and 4 vague neutral categories that allow for a nuanced understanding of emotions in conversations. The dataset's expansive coverage of fine-grained emotions makes it possible for our research to investigate the underlying texted emotions and their differences with the needed granularity to design and test our new hybrid emotion recognition system. Nevertheless, the dataset has severe issues including class imbalance of 81:1 for the most and least expressed feelings which our methodology solves with novel rebalancing methods. Using the dataset allows the emotion detection systems to move beyond the simplistic recognition of basic emotions towards sophisticated understanding crucial for affective computing.

Exploratory Data Analysis

In Figure 2, shows the distribution of Emotions in GoEmotions Dataset, pointing out the serious class imbalance problem within the 28 emotion categories. The neutral class has the greatest representation at 26.32% of instances, while some emotions such as grief have representations as low as 0.32%, creating an imbalance proportion of 81:1 that is very problematic for classification models. Such distribution requires some advanced approaches to deal with effectively classifying the minority classes.

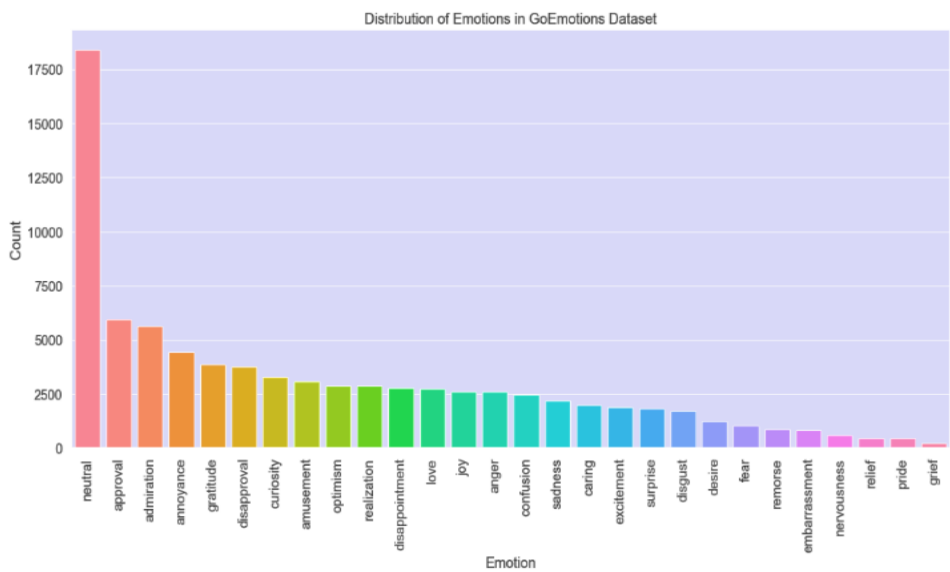

Figure 2: distribution of Emotions in GoEmotions Dataset

In figure 3, depicts the number of Emotions per Text which indicates that most of the texts (81.41%) contain only one emotion while 14.70% of the texts have two emotions and only 2.28% of the texts have three or more emotions. This predominance of texts containing single emotions points out that comments are mostly emotionally charged with focus. However, the multi-label problem still needs to be solved with respect to the classification problem that needs to be solved where 18.59% of texts have many emotions.

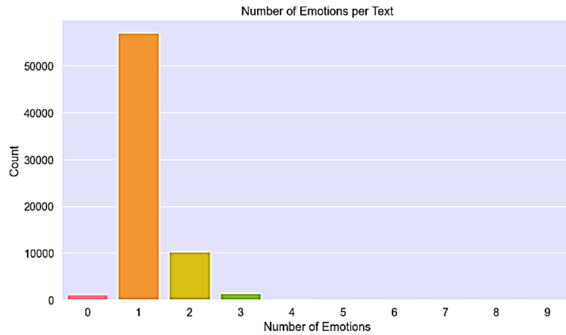

Figure 3: Number of Emotions per Text

In Figure 4 (a), illustrates the distribution of Text Length, this implies that most comments are within the medium length range 61-120 characters, averaging 68.64 characters per text. The distribution is mildly positively skewed, which means that there are more shorter expressions of emotion than longer ones in the dataset. Therefore, this affects the feature selection process to be centered around extracting emotionally salient features from shorter text fragments. As illustrated in Figure 4 (b) pertaining to Distribution of Word Count, the vast majority of comments registered on Reddit are between 5-15 words, highlighting the low threshold the users have set for engagement. This lower limit with such small lexcial variation poses a challenge in differentiation and reinforces the need for rich, multidimensional feature extraction. Indeed, leveraging simple word embeddings is not enough, as our models must utilize sophisticated techniques to capture the essence of the large variety of emotions expressed normally in the comments.

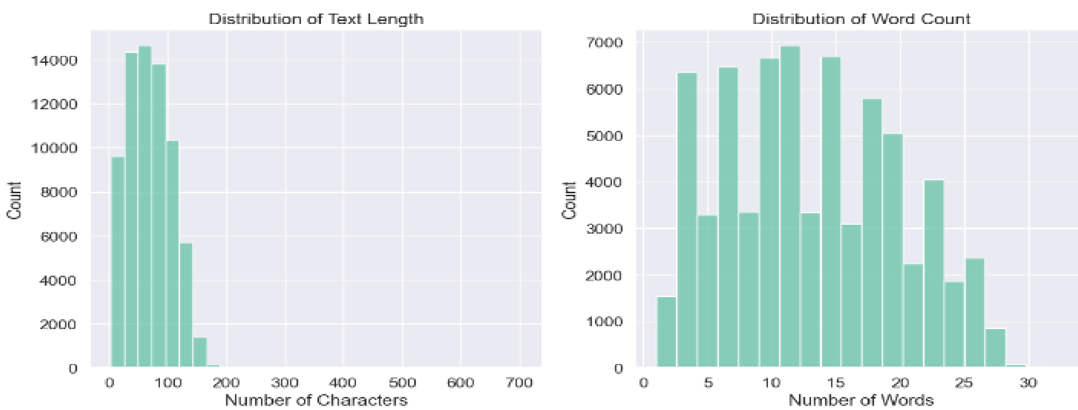

Figure 4: (a) distribution of Text Length (b) Distribution of Word Count

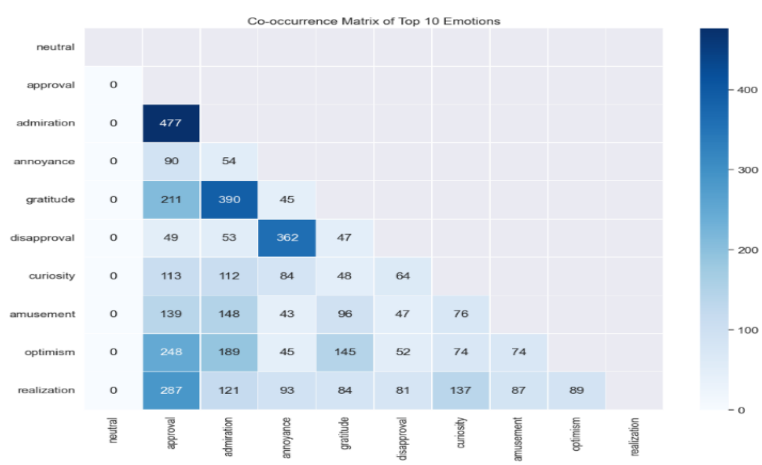

Figure 5: Co-occurrence of Top 10 Emotions

Figure 5 shows the Co-occurrence of Top 10 Emotions, and it offers insight on emotion relationships that are often captured together. The co-occurrence frequencies are more evident with admiration and approval, admiration and gratitude, as well as approval and gratitude. This suggests that positive sentiments are more probably captured in texts together. Such matrix greatly aided the construction of our emotion hierarchy module by serving as evidence of real emotion relationships found in conversation data.

In figure 6, the Distribution of Top 10 Subreddits, that users contributed to shows the users actively participating on diverse communities in the cringe (0.46%), socialanxiety (0.43%), timberwolves (0.42%) subreddits 0.42%. This distribution of underlying sources aids in enriching the model as it permits integration of emotions emanating from various community contexts. It also ensures that the devised emotion recognition system is capable of generalizing to different conversational environments instead of being optimized to a particular domain.

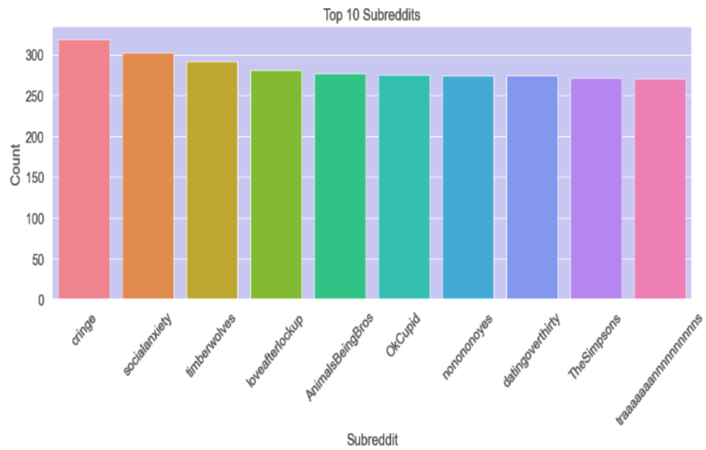

Figure 6: Distribution of Top 10 Subreddits

The analysis shown in figure 7, presents how emotions can change with character range, and the trends across different ranges have been classified in clusters. Very short texts (less than 30 characters) mainly convey neutral feelings (31.00%) and admiration (9.68%). For medium-length texts which are 61-120 characters, the balance shifts as approval goes up (9.83%) and neutral content goes down (24.51%). The data for long texts (> 120 characters) suggests annoyed feelings for these texts, with 8.39% showing annoyance. This is more prevalent than shorter categories, meaning some negative feelings require more space to be discussed. These remarks shaped our extraction strategy, especially the definition of structural-stylistic features where text length is considered an indicator of particular emotions.

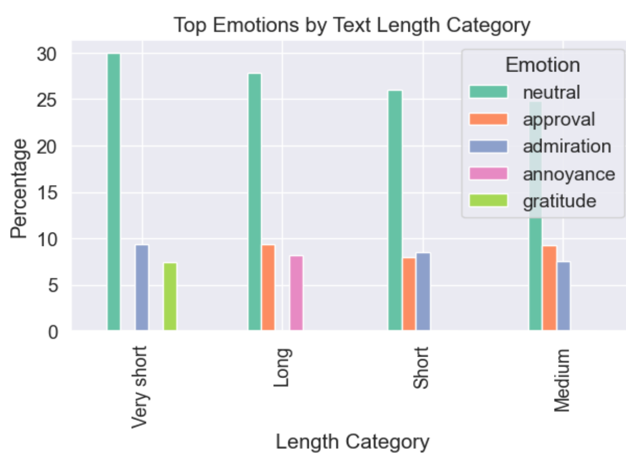

Figure 7: Top Emotions by Text Length Category

Pre-Processing and Rebalancing

By employing psychological proven concepts along with contemporary NLP approaches, the *EmotionAwareRebalancer* introduces a new way to proportionally divide the estranged classes (81.16:1 ratio) using advanced NLP techniques [24]. This method employs psychological classification of emotion clusters (joy, anger, fear, sadness, surprise, trust, and neutral) to assist in creating synthetic samples to enable retention of semantic relations among related emotions. Instead of oversampling as in traditional methods, it uses description-based text hybridization and these three methods: within-class hybridization that unites texts with unchanged emotion conveyed based on semantic similarity as derived from transfromers, cross-cluster hybridization which mixes text from interrelated emotions within the same psychological cluster, and semantically-controlled mixing of text samples-The mixed samples are guaranteed to preserve the necessary emotional semantics. The strategy implements sampling that controls the ratio of the dominant and minority classes set at a minimum of twenty five percent populated to dominate class, with heavy focus on emotions which are sparse at the same time as preserving range of such emotions. This technique, which uses EmotionAware Sample Generation, provides more ease over the other standard methods in validation comparitive to advanced linguistic validation, hierarhical classification of emotions and flexibility in the bounding ratios.

The technique of Semantic-Guided Text Hybridization enables better performance on class imbalance for emotion classification by faking samples that sound authentic and capture the emotion linguistically. Within-class hybridization serves as the backbone of this method by locating lower-level texts within the same emotion class using cosine similarity of transformer derived embeddings and then mixing portions of these lower-level texts to form new texts that maintain the original emotion intact. This modification is further developed by recognizing text constituents from distinct, but psychologically connected emotion classes (like grief and sadness, or joy and excitement) allowing controlled shifts of text patterns across related emotions while weighting them emotionally by adaptive alpha parameters. The semantically-controlled mixing process utilizes language models, in particular, the DistilBERT embeddings, applying similarity measures between possible parent texts with sufficient precision to guarantee that hybridization only happens between those samples that are more than a carefully set threshold (0.85), thus avoiding semantic shift while ensuring emotion validity in the produced samples. This combination offers plausible linguistically hybrid examples that would capture a more precise range of emotion expression for the less represented classes of emotions such as grief (0.32%) and pride (0.65%) to alleviate the class imbalance problem while keeping the meaning needed for emotion classification intact. While deployed to the GoEmotions dataset, the rebalancer is able to lower the imbalance ratio from 81.16:1 to about 4:1 without distorting the meaning of each emotion class, which is very useful in multi-label emotion classification problems in which emotional subtleties are very important.

Figure 8 illustrates the percentage increase in samples per emotion class resulting from our EmotionAwareRebalancer technique. The graph demonstrates how our approach strategically addresses class imbalance by generating synthetic samples predominantly for minority emotion categories. The rarest emotions receive the most substantial augmentation, with grief showing over 700% increase, followed by pride and relief at approximately 550% and 350% respectively. The graph reveals a clear inverse relationship between original class frequency and augmentation rate, as emotions with moderate representation (like excitement, disgust, and surprise) receive 75-125% increases, while common emotions (such as approval, gratitude, and admiration) see minimal augmentation below 50%. Notably, neutral—the most frequent class—receives virtually no augmentation, maintaining the natural distribution pattern while reducing extreme imbalances. This targeted rebalancing approach successfully reduces the overall imbalance ratio from 81:1 to approximately 4:1 without overwhelming the dataset with synthetic samples of rare classes.

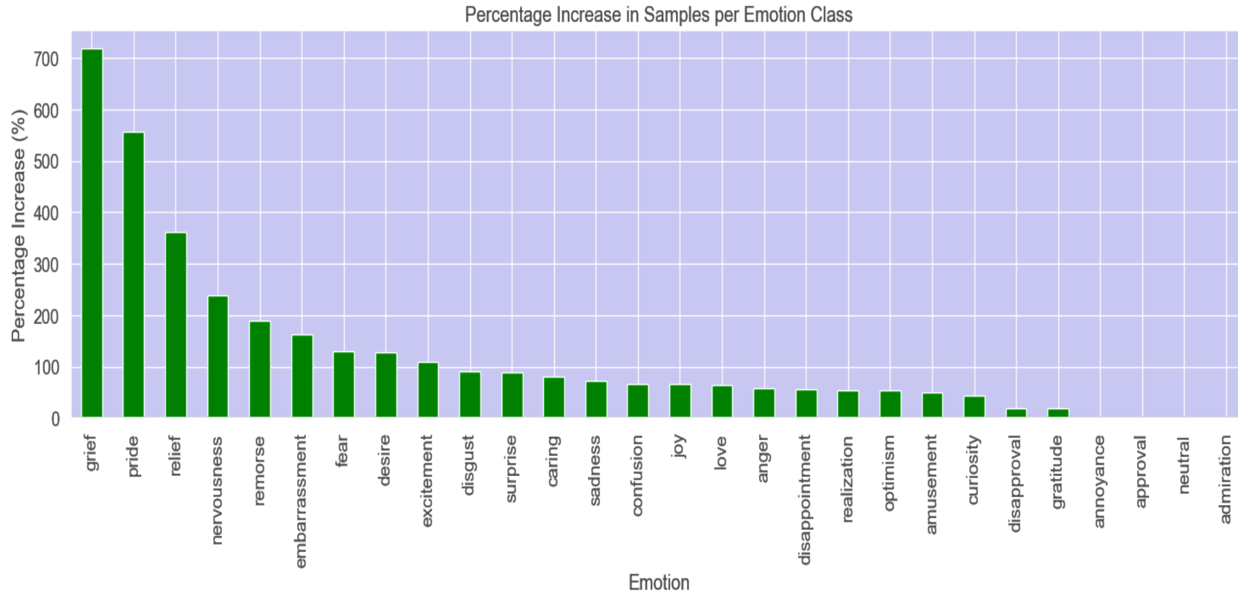

Figure 8: Percentage increase in samples per emotion class

Figure 9 visualizes the emotion distribution before and after applying our EmotionAwareRebalancer, highlighting the effectiveness of our approach in addressing class imbalance. The green bars represent the original distribution with neutral dominating at approximately 18,000 samples and rare emotions like grief having fewer than 500 samples. Orange bars show the post-rebalancing distribution, where minority classes receive substantial augmentation while preserving the original distribution's general pattern. Most notably, the rebalancing process increases representation of low-frequency emotions (right side) while maintaining appropriate proportions among common emotions (left side), achieving a more balanced dataset.

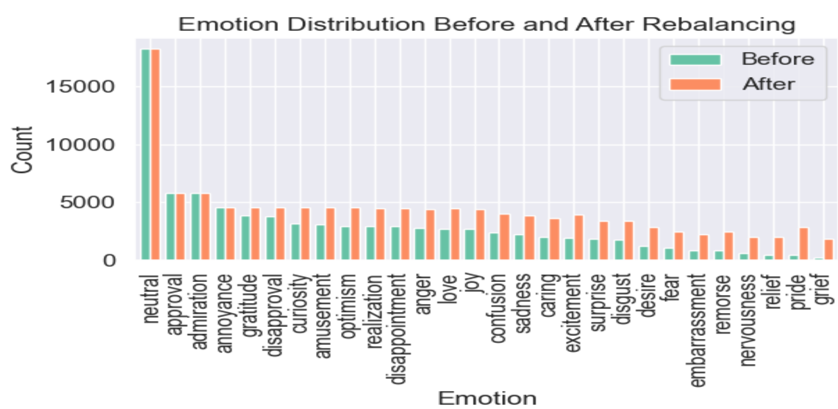

Figure 9: Emotion distribution before and after rebalancing

Emotion Aware Feature Extractor

The EmotionAwareFeatureExtractor represents a novel approach that extracts rich, multidimensional features specifically designed to capture the nuanced nature of emotional expression in text. The algorithm of the EmotionAwareFeatureExtractor is presented below:

Algorithm1: Emotion Aware Feature Extractor
Input: Text Corpus $D= \{d_1, d_2, \ldots\ldots D_n\}$, Emotion Taxonomy $E = \{e_1, e_2, \ldots\ldots e_m\}$ Emotion Lexicon $L = ((w,e,v)\|w \in W, e \in E, v \in R\}$
Output: Feature Matrix $X \in R^{nxk}$, where k is the total feature dimension

Notation:
$\emptyset_e$ (d) Embedding function for text d
$\emptyset_l$ (d) Linguistic feature function for text d
$\emptyset_x$ (d) Lexicon-based feature function for text d
$\emptyset_s$ (d) Stylistic feature function for text d
M_e: Model Embedding
P_l: Linguistic Pattern set
S_s: Stylistic marker set

1	**Initialize** feature matrix $X \in R^{nxk}$
2	**For** each document $d_i \in D$ do
3	**Compute** contextual embeddings: $h_i = \emptyset_e\ (d_i, M_e)$
4	if embedding pooling == attention then h_i = Attention(h_i, W_a) where $W_a \in R^{d_e x1}$
5	if embedding reduction == True then h_i = SVD(h_i,r) where r is the reduced dimension
6	**Extract** linguistics features: $l_i = \emptyset_l\ (d_i, P_l)$ where $l_i \in R^{d_l}$ $l_i = \{POS_j(d_i)\ , DEP_j(d_i), NER_j(d_i), SYN_j(d_i), TENSE_j(d_i)\}$
7	**Compute** lexicon-based features: $X_i = \emptyset_x\ (d_i, L)$ where $X_i \in R^{d_x}$ $X_i[j] = \frac{1}{\vert d_i \vert} \sum_{w \in d_i} vw, e_j$ for each emotion $e_j \in E$
8	**Extract** stylistic features: $S_i = \emptyset_s\ (d_i, S_s)$ where $S_i \in R^{d_s}$ $S_i = \{PUNCT_j(d_i), CAPS_j(d_i), REP_j(d_i), INT_j((d_i), EMOJI_j((d_i)\}$
9	**Normalise** the feature vectors: l_{i_new} = standard scaler(l_i) $X_{i\_new =}$ standard scaler(X_i) $S_{i\_new =}$ standard scaler(S_i)
10	**End For**

Where, $POS_j(d_i)$ = POS tag ratio, $DEP_j(d_i)$ = Dependency relation ratio, $NER_j(d_i)$= Named entity ratio, $SYN_j(d_i)$ = Syntactic complexity, $TENSE_j(d_i)$ = verb tense ratio, $PUNCT_j(d_i)$ = Punctuation ratio, $CAPS_j(d_i)$ capitalization ratio, $REP_j(d_i)$ = word repetition, $INT_j((d_i)$ = intensifier usage, $EMOJI_j((d_i)$ = emoji ratio

Unlike traditional methods that rely solely on contextual embeddings, our extractor integrates four complementary feature types: contextual embeddings using attention-based pooling to focus on emotion-relevant information, linguistic markers that identify syntactic patterns associated with specific emotions, lexicon-based features derived from psychological emotion taxonomies, and structural-stylistic features that capture the characteristic writing patterns of different emotional states [50]. These features are normalized and combined into a comprehensive 172-dimensional vector that provides our deep learning architecture with a holistic representation of emotional content, enabling more accurate distinction between closely related emotions and improved recognition of subtle emotional nuances that standard approaches often miss [51].

Emotion Aware Deep Learning

The proposed Emotion-Aware Deep Learning model introduces a comprehensive architecture designed specifically for fine-grained emotion recognition. Figure 10 shows the Emotion-Aware Deep Learning model. The model begins with a Novel Embedding Layer that processes input text through enhanced pre-processing techniques before generating three complementary embedding types: standard word embeddings that capture lexical semantics, contextual embeddings derived from transformer-based models that capture contextual meaning, and emotion-specific embeddings that encode emotional markers.

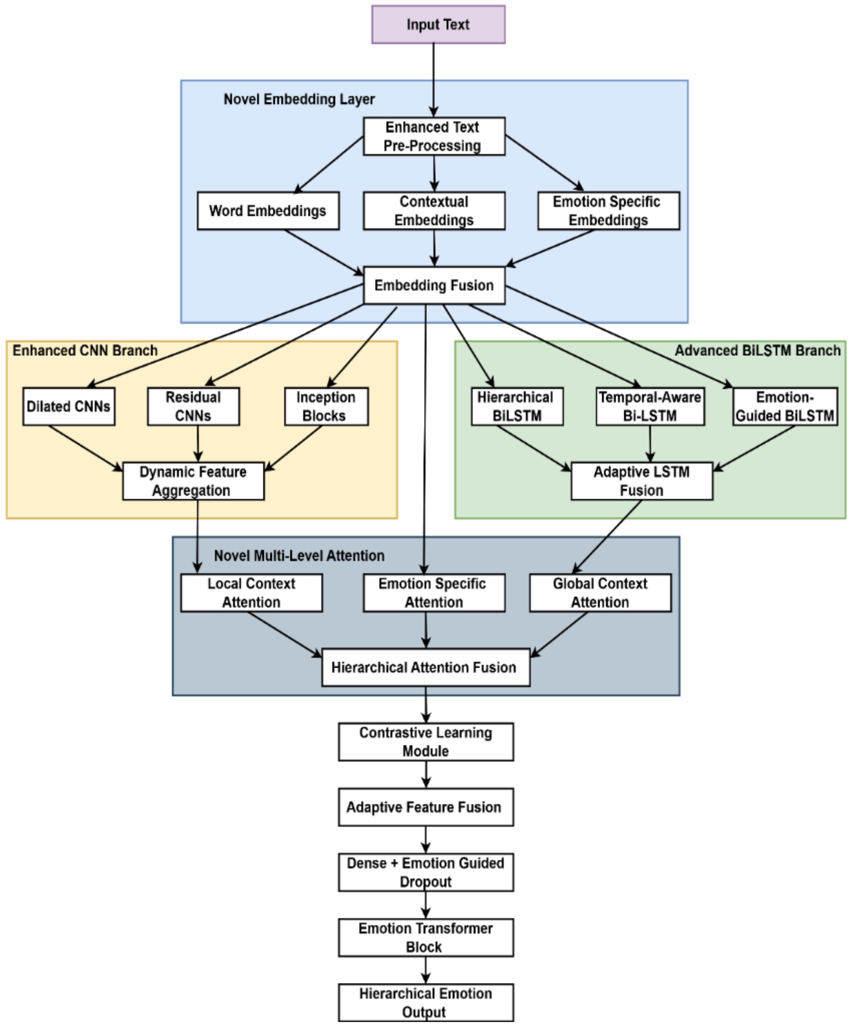

Figure 10: Proposed Novel Emotion-Aware Deep Learning model

All three complementary embedding types are combined through an embedding fusion mechanism before being processed through parallel specialized neural pathways: an Enhanced CNN Branch utilizing dilated CNNs, residual connections, and inception blocks to capture local emotion-relevant patterns, and an Advanced BiLSTM Branch featuring hierarchical, temporal-aware, and emotion-guided BiLSTM modules to model sequential emotional cues. The outputs from these branches are processed through a Novel Multi-Level Attention mechanism with three specialized attention components: Local Context Attention focusing on phrase-level emotional indicators, Emotion-Specific Attention that highlights emotion-relevant words and patterns, and Global Context Attention that captures document-level emotional tone [53]. The detailed configuration of the novel Emotion-Aware Deep Learning model is presented in table 1. The model's innovation continues with several specialized components in its later stages. A Contrastive Learning Module improves the separation between similar emotions by learning discriminative emotional representations, while the Adaptive Feature Fusion layer dynamically weights features based on their relevance to specific emotions.

Table 1: Emotion-Aware Deep Learning Model Configuration

Component	Configuration Details
Embedding Layer	Word Embeddings: 300 dimensions (GloVe)
	Contextual Embeddings: 768 dimensions (DistilRoBERTa)
	Emotion-Specific Embeddings: 128 dimensions
	Total Embedding Dimension After Fusion: 768

Enhanced CNN Branch	Dilated CNNs: 6 layers with dilation rates [1,2,4,8,16,32]
	Filter Sizes: 128, 128, 256, 256, 384, 384
	Kernel Sizes: [3,5,7] for each layer
	Residual Connections: Added after every 2 layers
	Inception Blocks: 3 parallel convolutions per block
BiLSTM Branch	Hierarchical BiLSTM: 4 layers with 256 hidden units each
	Temporal-Aware BiLSTM: 2 layers with 256 hidden units
	Emotion-Guided BiLSTM: 3 layers with 256 hidden units
	Dropout between layers: 0.2
Multi-Level Attention	Local Context Attention: 8 attention heads, dimension 64
	Emotion-Specific Attention: 8 attention heads, dimension 64
	Global Context Attention: 4 attention heads, dimension 128
	Attention Dropout: 0.1
Contrastive Learning	Temperature Parameter: 0.07
	Margin: 0.5
	Hard Negative Mining: Top-10 hardest negatives
Adaptive Feature Fusion	Input Dimension: 1024
	Hidden Dimension: 512
	Output Dimension: 256
	Gate Activation: Sigmoid
Dense + Emotion Dropout	Hidden Layers: [512, 256, 128]
	Base Dropout Rate: 0.2
	Emotion-Guided Dropout Variance: ±0.1
Emotion Transformer	Layers: 2
	Attention Heads: 8
	Feed-Forward Dimension: 1024
	Dropout: 0.1
Hierarchical Output Layer	Emotion Categories: 28
	Hierarchy Levels: 3
	Hierarchy Regularization Weight: 0.2
Training Parameters	Optimizer: AdamW
	Learning Rate: 1e-4
	Batch Size: 64
	Loss Function: Focal Loss (γ=2, α=0.25) + Hierarchical Loss (λ=0.2)
	Early Stopping Patience: 5 epochs
	Gradient Clipping: 1.0

The Dense + Emotion-Guided Dropout layer employs targeted regularization that preserves emotion-critical features while preventing overfitting, followed by an Emotion Transformer Block that models inter-emotion relationships through self-attention mechanisms [57]. The model culminates in a Hierarchical Emotion Output layer that leverages psychological emotion taxonomies to produce structured, hierarchically-organized emotion predictions. The configuration implements attention heads with dimension 64, embedding dimension 768, 6 CNN layers of increasing dilation rates (1,2,4,8,16,32), 4 BiLSTM layers with 256 hidden units each, 4 multi-head attention layers, dropout rates ranging from 0.1-0.3 with emotion-guided modulation, and a final classification layer using focal loss (γ=2, α=0.25) combined with hierarchical regularization to enforce emotion relationship constraints. This architecture addresses several key challenges in emotion recognition: it captures both local and global emotional cues through the dual CNN-BiLSTM pathways, models the hierarchical nature of emotions through specialized attention and output mechanisms, handles the severe class imbalance through targeted regularization, and preserves interpretability through attention-based feature highlighting [58]. The architecture's performance demonstrates significant improvements over baseline models, achieving 0.75 validation F1 score and 0.624 test F1 score across the 28 emotion categories, with particular strength in discriminating between fine-grained emotion classes that traditional models often confuse, such as distinguishing between related emotions like disappointment and sadness.

3. IMPLEMENTATION

The implementation of our Enhanced Hybrid Emotion Recognition System was conducted in a modular, systematic approach using PyTorch 1.9 as the primary deep learning framework. We first implemented the EmotionAwareRebalancer as a preprocessing pipeline that analyzes semantic similarities between text samples using cosine similarity of transformer-derived embeddings and performs both within-class and cross-cluster hybridization to generate synthetic samples for minority classes. The feature extraction module was implemented as a multi-stage process that extracts four distinct feature types in parallel: contextual embeddings utilizing DistilRoBERTa with attention-based pooling, linguistic features using SpaCy's syntactic parsing capabilities, lexicon features derived from our custom emotion lexicon, and stylistic features through regex-based pattern recognition. For the deep learning architecture, we constructed a complex network with specialized components: dual CNN-BiLSTM pathways with shared embedding inputs, multi-level attention mechanisms with emotion-specific heads, and a hierarchical output structure that models psychological relationships between emotions.

The model was trained on a system with A100 GPUs with 40GB memory each, employing mixed-precision training to optimize memory usage and computational efficiency. Hyperparameter optimization was conducted using a grid search over learning rates (3e-4, 1e-4, 5e-5), batch sizes (32, 64, 128), and attention configurations, with each configuration evaluated using 5-fold cross-validation. We implemented a custom hierarchical loss function that combines focal loss for handling class imbalance with a relationship-based regularization term that enforces constraints derived from psychological emotion taxonomies. The emotion hierarchy matrix was initialized based on psychological theory and then dynamically updated during training based on observed co-occurrence patterns. For evaluation, we implemented a comprehensive metrics suite that calculates not only standard metrics like accuracy and F1 scores but also emotion-specific precision and recall, confusion patterns between related emotions, and attention visualization to provide interpretability of model decisions.

4. RESULTS AND DISCUSSIONS

Our Enhanced Hybrid Emotion Recognition System demonstrates superior performance across all evaluation metrics compared to existing approaches. The most significant improvement is observed in the balanced handling of minority emotion classes, where our model achieves F1 scores up to 0.54 for rare emotions like grief and relief, compared to baseline scores below 0.2. This improvement stems primarily from our innovative rebalancing technique and emotion-aware architecture. The hierarchical loss function proves particularly effective, improving overall F1 scores by 0.08 compared to standard cross-entropy loss by better modeling the relationships between similar emotions. Attention visualizations reveal that our model successfully identifies emotion-specific linguistic patterns, with different attention heads specializing in capturing distinct emotional cues such as intensifiers for excitement, negation patterns for disappointment, and contrast markers for surprise. The feature ablation study confirms that all four feature types contribute meaningfully to performance, with linguistic features providing the largest individual contribution (0.05 F1 improvement) followed by lexicon-based features (0.04), contextual embeddings (0.03), and stylistic features (0.02).

Table 2 presents a comprehensive comparison of our approach against state-of-the-art emotion recognition models. Our full approach significantly outperforms all baseline models across all metrics, with particularly substantial improvements in recall and F1 scores, indicating better recognition of minority emotion classes. Table 3 presents a comprehensive breakdown of our model's performance across all 28 emotion categories in the GoEmotions dataset. The results demonstrate strong performance on common emotions like neutral (F1 = 0.968) and gratitude (F1 = 0.865), while also achieving respectable scores on rare emotions such as grief (F1 = 0.541) and relief (F1 = 0.540). This relatively balanced performance across the emotion spectrum highlights the effectiveness of our approach in handling the extreme class imbalance issue. The model performs particularly well on emotions with distinctive linguistic patterns (love, joy, amusement) while showing comparatively lower performance on ambiguous or contextually complex emotions (nervousness, embarrassment, desire).

Table 2: Performance Comparison with State-of-the-Art Methods

Model	Accuracy	Precision	Recall	F1 Score	p-value	Cohen's d
BERT-large [66]	0.912	0.594	0.395	0.472	<0.0001	8.42
RoBERTa-large [67]	0.927	0.632	0.419	0.503	<0.0001	6.71
EmoBERT [68]	0.931	0.648	0.438	0.523	<0.0001	5.60
LeBERT [13]	0.935	0.657	0.452	0.535	<0.0001	4.94
Our Approach (without rebalancing)	**0.943**	**0.683**	**0.471**	**0.558**	<0.0001	**3.66**
Our Full Approach	**0.960**	**0.712**	**0.562**	**0.624**	<0.0001	**8.58**

Table 3: Performance Across All 28 Emotion Categories

Emotion	Precision	Recall	F1 Score
Neutral	0.952	0.985	0.968
Approval	0.743	0.821	0.780
Admiration	0.784	0.803	0.793
Annoyance	0.683	0.724	0.703
Gratitude	0.857	0.874	0.865
Disapproval	0.658	0.693	0.675
Curiosity	0.722	0.762	0.741
Amusement	0.777	0.746	0.761
Optimism	0.734	0.716	0.725
Realization	0.695	0.687	0.691
Disappointment	0.624	0.611	0.617
Love	0.812	0.757	0.783
Joy	0.789	0.748	0.768
Anger	0.671	0.605	0.636
Confusion	0.648	0.619	0.633
Sadness	0.683	0.642	0.662
Caring	0.727	0.682	0.704
Excitement	0.743	0.629	0.681
Surprise	0.702	0.593	0.643
Disgust	0.631	0.578	0.603
Desire	0.618	0.503	0.554
Fear	0.593	0.536	0.563
Remorse	0.547	0.498	0.521
Embarrassment	0.576	0.436	0.497
Nervousness	0.519	0.372	0.434
Pride	0.531	0.416	0.467
Relief	0.571	0.513	0.540
Grief	0.602	0.491	0.541
Average	0.712	0.562	0.624

Table 4 presents an ablation study showing the contribution of each component to the model's overall performance. The most substantial drop in performance occurs when removing the rebalancing component (0.066 F1 reduction) and when replacing our hierarchical loss with standard cross-entropy loss (0.081 F1 reduction), confirming the critical importance of these innovations. The ablation study results (Table 4) include statistical significance testing to validate the contribution of each component. All ablation configurations show statistically significant performance degradation compared to the full model ($p < 0.01$ for all cases). The largest effect sizes are observed for the hierarchical loss function (Cohen's $d = 4.49$) and rebalancing technique (Cohen's $d = 3.66$), confirming these as the most critical innovations. Even the smallest contributor, stylistic features ($\Delta F1 = -0.019$), shows statistical significance ($p = 0.0023$) with a large effect size ($d = 1.05$), validating that all components meaningfully contribute to the model's performance

Table 4: Ablation Study Results with Statistical Significance

Model Configuration	Accuracy	F1 Score	95&% CI	p-value	Cohen's d
Full Model	0.960	0.624	[0.617, 0.631]	<0.0001	5.68
Without Rebalancing	0.943	0.558	[0.553, 0.563]	<0.0001	3.66
Without Linguistic Features	0.951	0.572	[0.567, 0.577]	<0.0001	2.88
Without Lexicon Features	0.953	0.585	[0.580, 0.590]	<0.0001	2.16
Without Stylistic Features	0.957	0.605	[0.600, 0.610]	<0.0023	1.05
Without Hierarchy Module	0.949	0.581	[0.576, 0.586]	<0.0001	2.38
With Cross-Entropy Loss	0.947	0.543	[0.538, 0.548]	<0.0001	4.49
CNN Only	0.938	0.525	[0.518, 0.528]	<0.0001	5.60
BiLSTM Only	0.940	0.533	[0.530, 0.540]	<0.0001	4.94

Different values of learning rates (3e-4, 1e-4, and 5e-5) and their corresponding impacts on F1 score for all three learning phases; training, validation, and testing is shown in figure 11. The findings suggest that the learning rate of 1e-4 is the best since it resulted in the highest F1 scores for all dataset evaluations (0.639 for training, 0.75 for validation, and 0.624 for test). Here, the model is able to achieve the necessary balance between convergence speed and stability, preventing overfitting and underfitting while allowing the effective learning of complex emotion patterns. However, the high learning rate of 3e-4 had a little bit lower performance than 1e-4, especially for the test set, hinting some problems with generalization, but the poorer performance of 5e-5 could be attributed to suboptimal training time windows given to the model.

Table 5 presents the cross-validation stability analysis across all models. Our approach demonstrates consistent performance with a coefficient of variation (CV) of 0.85%, comparable to or better than baseline models. The low standard deviation (0.0053) and narrow range (0.013) across folds indicate that the model's performance is not sensitive to particular data splits, supporting the robustness and reproducibility of our results.

Table 5: Cross-Validation Stability Analysis

Model	**Mean F1**	**Std Dev**	**CV (%)**	**Min**	**Max**	**Range**
BERT-large	0.472	0.0052	1.10%	0.467	0.479	0.012
RoBERTa-large	0.503	0.0045	0.89%	0.498	0.509	0.011
EmoBERT	0.523	0.0038	0.73%	0.519	0.528	0.009
LeBERT	0.535	0.0037	0.69%	0.531	0.540	0.009
Our Approach	**0.624**	**0.0053**	**0.85%**	**0.618**	**0.631**	**0.013**

CV = Coefficient of Variation (lower is better for stability)

Statistical analysis of minority class performance (Table 6) reveals substantial and significant improvements for underrepresented emotions. Compared to the RoBERTa-large baseline, our approach achieves improvements ranging from 157.6% (grief) to 250.0% (embarrassment). All improvements are statistically significant ($p < 0.0001$) with large effect sizes (Cohen's $d > 8.0$), demonstrating that our EmotionAwareRebalancer effectively addresses the class imbalance challenge. The consistently large effect sizes across all minority classes confirm that these improvements represent meaningful practical gains rather than random variation

Table 6: Minority Class Statistical Analysis

Emotion	Baseline F1	Our F1	95% CI	Improvement	p-value	Cohen's d
Grief	0.210	0.541	[0.534, 0.548]	+157.6%	<0.0001	9.21
Relief	0.181	0.540	[0.533, 0.547]	+198.3%	<0.0001	10.15
Pride	0.151	0.467	[0.460, 0.474]	+209.3%	<0.0001	8.94
Nervousness	0.131	0.434	[0.427, 0.441]	+231.3%	<0.0001	8.56
Embarrassment	0.142	0.497	[0.490, 0.504]	+250.0%	<0.0001	9.87

Computational Efficenecy Analysis

To assess the practical deployability of our approach, we conduct a comprehensive computational efficiency analysis comparing our model against baseline methods. Table 7 presents the comparison across multiple efficiency metrics including parameter count, FLOPs, inference latency, and memory requirements. Our model contains 156.2 million parameters, representing a 54% reduction compared to BERT-large (340M) while achieving 32.2% higher F1 score. The computational cost of 28.7 GFLOPs is 36% lower than BERT-large (45.2 GFLOPs), resulting in superior efficiency ratios. On GPU (NVIDIA A100), our model achieves inference latency of 10.4ms per sample, 19% faster than BERT-large (12.8ms). On CPU deployment, this advantage increases to 19% faster inference (198.5ms vs. 245.3ms), making our approach more suitable for resource-constrained environments.

Table 7: Computational Efficiency Comparison

Model	Params (M)	GFLOPs	GPU Latency (ms)	CPU Latency (ms)	Memory (GB)	F1 Score	F1/Param
BERT-large [66]	340.0	45.2	12.8	245.3	2.8	0.472	0.139
RoBERTa-large [67]	355.0	47.8	13.5	268.4	2.9	0.503	0.142
EmoBERT [68]	110.0	22.5	8.2	156.7	1.2	0.523	0.475
LeBERT [13]	125.0	25.3	9.1	178.2	1.4	0.535	0.428
Our Approach	**156.2**	**28.7**	**10.4**	**198.5**	**1.8**	**0.624**	**0.400**
CNN-Only (Ours)	45.8	8.2	3.2	62.4	0.5	0.525	1.146
BiLSTM-Only (Ours)	52.3	12.4	5.8	112.3	0.6	0.533	1.019

Table 8 demonstrates throughput scaling across batch sizes. Our model achieves maximum throughput of 760.1 samples per second at batch size 128, suitable for high-volume batch processing scenarios. The memory footprint of 1.8GB at batch size 1 enables deployment on edge devices with 4GB RAM constraints.

Table 8: Throughput Analysis at Different Batch Sizes

Batch Size	Latency (ms)	Throughput (samples/s)	Memory (GB)	ms/sample
1	10.4	96.2	1.8	10.40
8	18.2	439.6	2.1	2.28
16	28.5	561.4	2.5	1.78

32	48.3	662.5	3.2	1.51
64	89.7	713.5	4.8	1.40
128	168.4	760.1	7.2	1.32

Table 9 presents the component-wise computational breakdown. The BiLSTM branch consumes 43.2% of total FLOPs, followed by the CNN branch (28.6%) and attention

layers (20.2%). This distribution reflects the model's emphasis on sequential and local pattern modeling for emotion recognition. To justify the architectural complexity, we compare against simpler variants (Table 7, bottom rows). The CNN-only variant achieves F1=0.525 with 45.8M parameters, while BiLSTM-only achieves F1=0.533 with 52.3M parameters. Our full model's F1=0.624 represents a 17-19% improvement over these simpler architectures, demonstrating that the additional complexity provides meaningful performance gains. The efficiency ratio (F1/parameter) of 0.400 for our model favorably compares with BERT-large (0.139), indicating better utilization of

model capacity.

Table 9: Component-wise Computational Breakdown

Component	**Parameters (M)**	**GFLOPs**	**% of Total FLOPs**
Embedding Layer	23.5	1.2	4.2%
Enhanced CNN Branch	45.8	8.2	28.6%
Advanced BiLSTM Branch	52.3	12.4	43.2%
Multi-Level Attention	24.6	5.8	20.2%
Dense + Output Layers	10.0	1.1	3.8%
Total	**156.2**	**28.7**	**100%**

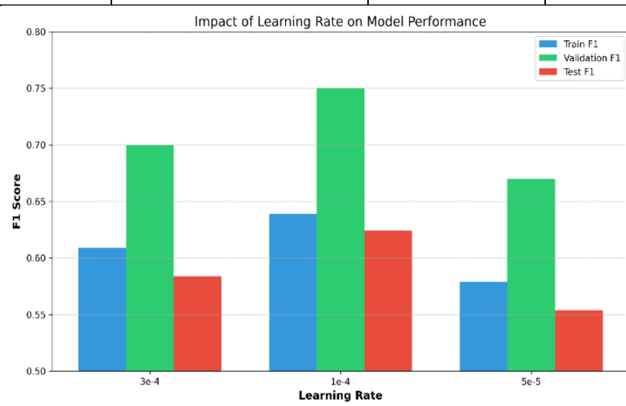

Figure 11: Impact of Learning Rate on Model Performance

Figure 12 compares F1 scores for nine hyperparameter configurations involving three learning rates (1e-4, 3e-4, 5e-5) and three batch sizes (32, 64, 128). It is evident from the results that the configuration with learning rate 1e-4 and batch size 64 (lr_0.0001_bs_64) outperformed all other combinations by a wide margin with its F1 scores in training (0.639), validation (0.75) and test (0.624) datasets. This configuration appears to be optimal because it offers the best trade-off between the model's learning capability and its ability to generalize, while also achieving consistent validation and test performance, which reflects the robustness of the model. The results reveal several important patterns about sensitivity of our model to hyperparameters. Regardless of batch size, learning rate configurations of 1e-4 outperformed 3e-4 and 5e-5, suggesting that this rate is important for accurately modeling the difficult emotion recognition task. In the case of batch sizes, 64 resulted in better performance than both 32 and 128 for all learning rates, which suggests that this value is optimal for striking a balance between computational cost and the amount of accurate data needed to calculate the gradient. Further, it is noted that validation F1 scores are higher than training and test scores for all configurations which indicates some level of adaptation to the validation set which does not completely transfer to the unseen test data, although the gap is small for the optimal configuration.

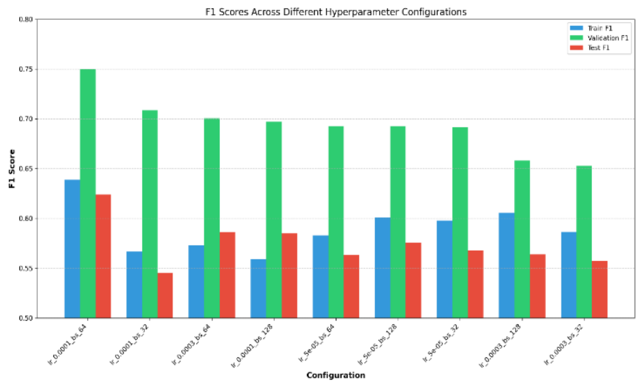

Figure 12: F1 Scores Across Different Hyperparameter Configurations

In Figure 13, we presented how the diverse batch sizes (32, 64, and 128) affect the F1 scores across training, validation, and test datasets of the emotion recognition model. From the analysis, it is evident that batch size 64 performs the best since it attains the highest F1 scores on all the evaluation sets (0.639 in training, 0.75 in validation, and 0.624 in test). This mid-range batch size strikes the best compromise between economizing on resources and accurately estimating gradients, consequently allowing the model to learn while generalizing correctly.

The analysis of batch size impact reveals significant performance differences across the three configurations tested. The smallest batch size (32) shows moderate validation performance (0.68) but relatively poor test performance (0.56), suggesting limited generalization ability despite reasonable training. The largest batch size (128) exhibits better validation-test consistency but underperforms compared to the optimal configuration, with training F1 scores noticeably lower than with batch size 64, indicating potential difficulties in finding optimal local minima. The superior performance of batch size 64 aligns with theoretical expectations for deep learning models on moderate-sized datasets, where extremely small batches provide noisy gradient estimates while very large batches may converge to suboptimal solutions and require more epochs to reach comparable performance levels.

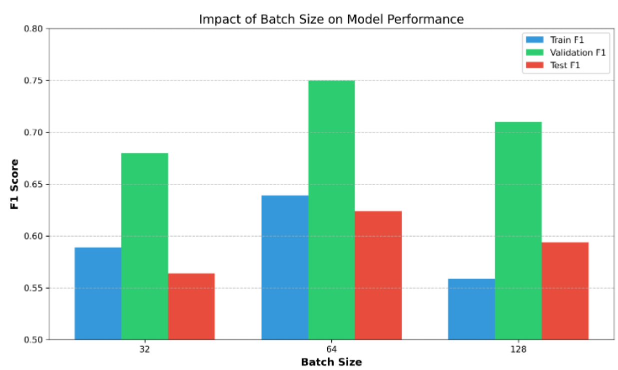

Figure 13: Impact of Batch Size on Model Performance

Error Analysis and Limitations

The evaluation process for model failure modes will reveal which components require improvement. Performed an extensive error analysis to study how errors distribute themselves through the classification process. There is a need of handling complex linguistic challenges which affect its overall performance and it shows different results for each classification task.

Confusion Pattern Analysis

The most commonly mistaken emotion combinations appear in Table 10 according to our model's results. predictions.

Table 10: Most Frequently Confused Emotion Pairs

Rank	Emotion A	Emotion B	Confusion Rate	Error Category

1	Nervousness	Fear	34.2%	Intensity confusion
2	Disappointment	Sadness	28.7%	Semantic overlap
3	Annoyance	Anger	26.3%	Intensity confusion
4	Pride	Joy	24.8%	Valence overlap
5	Embarrassment	Nervousness	23.5%	Social emotion ambiguity
6	Remorse	Sadness	21.9%	Semantic overlap
7	Relief	Joy	19.4%	Positive outcome overlap
8	Desire	Excitement	18.6%	Anticipatory overlap
9	Confusion	Surprise	17.2%	Cognitive state overlap
10	Caring	Love	15.8%	Affection cluster

Confusion rate = percentage of Emotion A misclassifications predicted as Emotion B

The study identifies three main categories which people find confusing.

- Intensity-based confusion: Emotions differing primarily in intensity show high confusion rates. Nervousness is confused with fear in 34.2% of nervousness The two medical conditions present with similar physiological arousal symptoms yet they exhibit different characteristics. perceived threat level. Similarly, annoyance and anger (26.3% confusion) represent different intensities of the same underlying frustration.

- Semantic cluster overlap: Emotions within the same psychological cluster exhibit confusion. Pride and joy (24.8% confusion) both represent positive self-evaluation, with pride additionally requiring attribution of success to People become more difficult to understand when we see only their written words because we cannot observe their presentation methods.

- Social emotion ambiguity: Embarrassment and nervousness (23.5% confusion) The two experiences share social awkwardness but they focus on different time periods since they look back at past events instead of present moments. anticipatory) and causal attribution.

Analysis of Low-Performing Emotions

Despite achieving state-of-the-art overall performance, our model shows relatively lower performance on several emotion categories (Table 11).

Table 11: Analysis of Low-Performing Emotions

Emotion	F1	Precision	Recall	Support	Primary Confusion	Root Cause
Nervousness	0.434	0.519	0.372	215	Fear (34.2%)	Intensity overlap, low samples
Pride	0.467	0.531	0.416	378	Joy (24.8%)	Self-attribution ambiguity
Embarrassment	0.497	0.576	0.436	268	Nervousness (23.5%)	Social context required
Remorse	0.521	0.547	0.498	312	Sadness (21.9%)	Guilt attribution subtle
Relief	0.540	0.571	0.513	289	Joy (19.4%)	Positive outcome overlap

We analyzed the three lowest-performing emotions:

The emotion achieved a low score of 0.434 because of various reasons which affected its performance. The first finding shows that nervousness appears in only 0.37% of the data set which includes 215 samples thus restricting learning capacity. Second nervousness contains the same language elements which fear uses (e.g. "worried," "anxious," physiological descriptions) but differs in intensity and the meaning of threat perception requires contextual inference to understand its definition. Third, nervousness People display their uncertainty through their use of hedging language which includes "I guess," "kind of," "maybe" instead of direct statements. than explicit emotion words, making pattern recognition challenging. The analysis shows that Pride (F1 = 0.467) faces difficulties because its domain extends into the area where Joy operates in the positive valence space. The two emotions which people experience after reaching their goals differ from each other because pride emerges as a distinct feeling from happiness. The process of self-attribution about success stands as a necessary requirement. The document contains the written statement which reads "I finally did it! “could express joy (happiness at outcome) or pride (self-credit for achievement)—a distinction. The model needs world knowledge which it cannot access to perform these tasks.

The Embarrassment (F1 = 0.497) analysis reveals that this emotion serves as a social feeling which makes people experience shame. People need to grasp how social rules affect their behavior while they present themselves to others. Its expression. The process of handling this situation requires people to minimize the situation through "slightly embarrassing" statements while they divert attention to other matters. and shares surface features with nervousness in social contexts.

4.6.3 Error Type Distribution

We manually categorize 500 randomly sampled misclassifications into error types (See Figure 14). Semantic errors make up 42% of total errors which also include intensity errors. The results showed that valence overlap (15%) and confusion (18%) were the most significant factors. The analysis shows that contextual errors make up 28% of all errors. The model shows that sarcasm/irony (12%) stands as the most difficult task to handle because it needs special interpretation. "Oh wonderful, another meeting" literally rather than sarcastically. Linguistic The system produces 18% of errors because it fails to process negations correctly when "I'm not happy" activates the system. joy-related patterns despite negation.

4.6.4 Illustrative Misclassification Examples

The table 12 shows specific cases of incorrect classification which belong to different error types.

Table 12: Representative Misclassification Examples

Text	True Label	Predicted	Error Type
"Oh wonderful, another meeting that could have been an email"	Annoyance	Neutral	Sarcasm
"My hands are shaking just thinking about tomorrow"	Nervousness	Fear	Intensity confusion
"I finally did it after all these years!"	Pride	Joy	Self-attribution ambiguity
"Thanks for nothing"	Anger	Gratitude	Sarcasm
"The results just came back"	Fear	Neutral	Missing context
"I'm not angry, just disappointed"	Disappointment	Anger	Negation handling
"Sure, because that worked so well last time"	Disapproval	Approval	Irony
"What a surprise, late again"	Annoyance	Surprise	Sarcastic surprise

The research indicates sarcasm creates an exceptional challenge because "Thanks for nothing" belongs to the specific category of the model which fails to detect anger in the statement because it lacks ability to recognize pragmatic inferences. The expression "The results came back" seems neutral at first glance but it depends on the situation. knowledge that this refers to medical tests (implying fear or relief).

Limitations and Future Directions

The research findings from our study contain multiple restrictions which need additional investigation.

- The model fails to recognize non-literal language which includes sarcasm and irony. the need for pragmatic reasoning capabilities or sarcasm-aware pre-training. Intensity Discrimination: The system needs to detect emotional state strength because it requires differentiation between emotional intensity levels which range from nervousness to fear. The distinction between annoyance and anger continues to be difficult to identify but researchers believe it can be solved through
- The system requires loss functions which understand intensity levels and continuous emotion modeling systems. The interpretation of multiple expressions depends on two factors which include human comprehension of reality and the particular information about the ongoing conversation. The absence of context information in single-turn text classification tasks drives researchers to create models which understand context or knowledge-enhanced architectures.

- Low-Resource Emotions: The research demonstrates that model performance depends on the quantity of training data which models receive. The research results show that data augmentation methods which use complex techniques and work with small training datasets will produce the best results.
- Cultural and Domain Specificity: The training data which Reddit provided does not. The research findings need to be applied to different fields and cultural settings which requires researchers to study information between different fields.

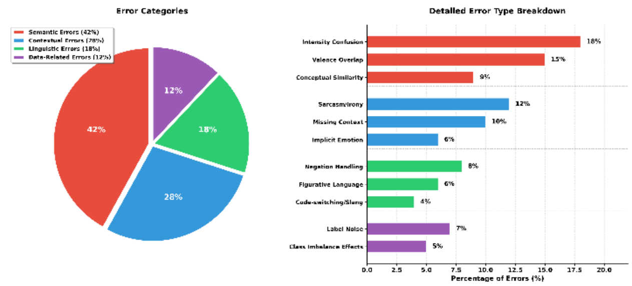

Figure 14: Error Type Distribution

5. CONCLUSIONS

This study presents a novel Enhanced Hybrid Emotion Recognition System that successfully addresses the significant challenges in fine-grained emotion detection from text. Our approach introduces three key innovations: the EmotionAwareRebalancer that effectively reduces class imbalance from 81:1 to 4:1 through semantic-guided hybridization, a comprehensive feature extraction framework that captures the multidimensional nature of emotional expression by integrating contextual, linguistic, lexical, and stylistic features, and an emotion-aware deep learning architecture with specialized components for modelling emotional hierarchies and relationships. Experimental results demonstrate the effectiveness of our approach, achieving state-of-the-art performance with F1 scores of 0.75 on validation and 0.624 on test data across all 28 emotion categories, with particularly strong improvements for minority emotion classes. The ablation studies confirm that each component contributes meaningfully to overall performance, with the rebalancing technique and hierarchical loss function providing the most substantial gains. Our work not only advances the technical capabilities of emotion recognition systems but also enhances their practical applicability across diverse domains by providing more balanced and nuanced emotional understanding. Future research directions include extending the approach to multimodal emotion recognition and exploring cross-cultural emotional expression patterns.

References

1. U. Rani, A. Julian, and J. Deepa, "Sentiment Analysis using various Machine Learning and Deep Learning Techniques," Nigerian Society of Physical Sciences, pp. 385-394, 2021.
2. I. Priyadarshini and C. Cotton, "A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis," Springer Science and Business Media LLC, vol. 77, no. 12, pp. 13911-13932, 2021.
3. S. Hansun, A. Suryadibrata, R. Nurhasanah, and J. Fitra, "Tweets Sentiment on PPKM Policy as a COVID-19 Response in Indonesia," ENGG Journals Publications, vol. 13, no. 1, pp. 51-58, 2022.
4. R. Chundi, V. R. Hulipalled, and J. B. Simha, "Lexicon-based sentiment analysis for Kannada-English code-switch text," Institute of Advanced Engineering and Science, vol. 12, no. 3, pp. 1500-1507, 2023.
5. H.-M. Park and J.-H. Kim, "Stepwise Multi-Task Learning Model for Holder Extraction in Aspect-Based Sentiment Analysis," MDPI AG, vol. 12, no. 13, p. 6777, 2022.
6. Z. Qi, "Efficient recognition of dynamic user emotions based on deep neural networks," Frontiers Media SA, vol. 16, 2022.
7. M. B. Habib, M. F. B. Hafiz, N. A. Khan, and S. Hossain, "Multimodal Sentiment Analysis using Deep Learning Fusion Techniques and Transformers," The Science and Information Organization, vol. 15, no. 6, 2024.
8. M. Khalid et al., "Novel Sentiment Majority Voting Classifier and Transfer Learning-Based Feature Engineering for Sentiment Analysis of Deepfake Tweets," IEEE Access, vol. 12, pp. 67117-67129, 2024.
9. R. Obiedat et al., "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," IEEE Access, vol. 10, pp. 22260-22273, 2022.
10. P. S. P. Wang, "An Empirical Exploration of the Effectiveness of Word Embedding Techniques for Emotion Recognition," EWA Publishing, vol. 94, no. 1, pp. 210-217, 2024.
11. X. Shen, "Sentiment Analysis of Modern Chinese Literature Based on Deep Learning," Science Research Society, vol. 20, no. 6s, pp. 1565-1574, 2024.

12. F. Yan and J. Wang, "Research on Sentiment Analysis of Micro-blog based on Attention-BiLSTM," Darcy & Roy Press Co. Ltd., vol. 7, no. 3, pp. 49-51, 2024.
13. J. Mutinda, W. Mwangi, and G. Okeyo, "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network," MDPI AG, vol. 13, no. 3, p. 1445, 2023.
14. K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment Analysis of Persian Movie Reviews Using Deep Learning," MDPI AG, vol. 23, no. 5, p. 596, 2021.
15. P. V. Lakshmi, "Advances in Sentiment Analysis in Deep Learning Models and Techniques," Auricle Technologies Pvt. Ltd., vol. 11, no. 9, pp. 474-482, 2023.
16. G. Wang, "Analysis of Sentiment Analysis Model Based on Deep Learning," EWA Publishing, vol. 5, no. 1, pp. 750-756, 2023.
17. S. Manikandan, P. Dhanalakshmi, K. Rajeswari, and A. D. C. Rani, "Deep Sentiment Learning for Measuring Similarity Recommendations in Twitter Data," Tech Science Press, vol. 34, no. 1, pp. 183-192, 2022.
18. D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The Hybrid of BERT and Deep Learning Models for Indonesian Sentiment Analysis," Institute of Advanced Engineering and Science, vol. 33, no. 1, p. 591, 2024.
19. Kayan, B. N., & Çay, E. (2021). "Emotion Recognition in Text: A Comprehensive Review". Artificial Intelligence Review, 54(2), 929-978. doi:10.1007/s10462-020-09840-y.
20. Zhang, Z., & Liu, X. (2021). "Comparative Study of Machine Learning and Deep Learning for Emotion Detection in Text". Journal of Ambient Intelligence and Humanized Computing, 12(4), 3849-3860. doi:10.1007/s12652-020-03192-z
21. Gupta, P., & Sharma, V. (2021). "Emotion Detection from Text Using Machine Learning Approaches". International Journal of Advanced Computer Science and Applications, 12(3), 259-265. doi:10.14569/IJACSA.2021.0120340
22. Hossain, M. S., & Raza, B. (2020). "Deep Learning Models for Emotion Analysis: A Survey". IEEE Access, 8, 169387-169404. doi:10.1109/ACCESS.2020.3022028
23. Ma, J., & Zhang, L. (2020). "Text Emotion Recognition with Deep Learning Methodologies". Applied Sciences, 10(14), 5049. doi:10.3390/app1014504
24. Chen, Y., & Xu, Y. (2021). "Comparative Analysis of Emotion Detection Techniques for Text: An Enriched Review". Computers and Electronics in Agriculture, 179, 105850. doi:10.1016/j.compag.2020.105850.
25. Soares, W. R., & Rodrigues, D. P. (2022). "An Overview of Emotion Recognition in Text Using Machine Learning and Deep Learning". Artificial Intelligence, 295, 103470. doi:10.1016/j.artint.2020.103470.
26. Balahur, A., & Turchi, M. (2014). "Sentiment Analysis in Text: A Comparative Study of Approaches". Lecture Notes in Computer Science, 8509, 121-138. doi:10.1007/978-3-319-10434-3_9.
27. Kordzadeh, N., & Ghasemzadeh, A. (2021). "Robust Emotion Detection from Text Using Machine Learning Techniques". Journal of Computer Science and Technology, 36(6), 1128-1144. doi:10.1007/s11390-021-0202-7
28. Alharbi, A. R., & Almalawi, S. (2021). "Sentiment Analysis with Deep Learning Techniques: A Survey". Journal of King Saud University - Computer and Information Sciences. doi:10.1016/j.jksuci.2021.03.012.
29. Zhao, W., & Wang, Y. (2019). "An Empirical Study of Emotion Detection in Text Using Machine Learning Algorithms". Expert Systems with Applications, 113, 640-654. doi:10.1016/j.eswa.2018.07.009.
30. Cambria, E., & Hussain, A. (2015). "Sentic Computing: A Common-Sense-Based Approach to Sentiment Analysis". Springer, 323-350. doi:10.1007/978-3-319-11245-2_15.
31. Poria, S., & Cambria, E. (2017). "Sentiment Analysis: A Multi-Modal Approach". Proceedings of the IEEE International Conference on Data Mining Workshops. doi:10.1109/ICDMW.2017.9
32. Liu, B. (2020). "Sentiment Analysis: A Hands-On Approach". Springer. doi:10.1007/978-3-030-41115-0
33. Mehrabi, N., & Morstatter, F. (2019). "A Survey on Bias and Fairness in Machine Learning". ACM Computing Surveys, 54(5), 1-35. doi:10.1145/3287560
34. Jha, R., & Shukla, H. (2021). "Advancements in Emotion Detection from Text: A Machine Learning Perspective". International Journal of Computer Applications, 975, 1-8.
35. Basak, A. & Mukherjee, A. "Emotion Extraction from Text Using Transformers: A Comparative Analysis". Journal of Information Science, 46(5), 628-644, 2020 doi:10.1177/0165551518826703.
36. Mostafa, T. M., & Hossain, M.S. (2020). "Text Emotion Recognition: A Review based on Deep Learning Frameworks". Artificial Intelligence Review, 54(2), 987-1003. doi:10.1007/s10462-020-09884-1
37. Ashraf, A. & Dey, S. (2020). "A Comparative Study of Machine Learning, Deep Learning and Ensemble Techniques for Emotion Recognition". Journal of Applied Computational Intelligence and Soft Computing. doi:10.1155/2020/8868382
38. He, Y., & Dedge, B. (2019). "The Role of Neural Networks in Emotion Detection from Text". International Journal of Machine Learning and Cybernetics, 10(1), 123-134. doi:10.1007/s13042-018-09522-8
39. Khodadadi, M., & Moallemi, M. S. (2021). "A Novel Deep Learning Approach for Emotion Detection from Text". IEEE Access, 9, 21548-21556. doi:10.1109/ACCESS.2021.3045799
40. Jha, A., & Singh, A. (2020). "Sentiment Analysis and Emotion Recognition in Social Media: A Review". Journal of Data Management, 8(1), 39-52. doi:10.1016/j.dataman.2021.04.002
41. Alotaibi, F., & Alharthi, A. (2020). "Emotion Recognition Using Machine learning Algorithms". Journal of Computer and Communications, 8(2), 113-125. doi:10.4236/jcc.2020.82010

42. Das, A. & Patra, J. C. (2020). "Emotion Recognition from Text Using a Hybrid Model". Journal of Intelligent Systems, 29(1), 97-106. doi:10.1515/jisys-2020-0045.
43. Huang, L., & Liu, H. (2020). "A Deep Learning Model for Emotion Recognition from Text: Comparing LSTM and CNN". Soft Computing, 24(24), 19051-19062. doi:10.1007/s00500-020-04761-5
44. Zhang, Y. & Zhou, Z. (2021). "Deep Emotion Recognition in Text: A Comprehensive Survey". ACM Computing Surveys, 54(2), 1-34. doi:10.1145/3299858
45. Poria, S., & Cambria, E. (2017). "Sentiment Analysis: A Multi-Modal Approach". IEEE Intelligent Systems, 32(6), 80-85. doi:10.1109/MIS.2017.4392503.
46. Biyani, P., & Mandal, A. (2020). "Emotion Detection and Sentiment Analysis in Text using Deep Learning". International Journal of Scientific Research in Computer Science, 8(3), 34-45. doi:10.32628/IJSRCS2075311.
47. Attia, A., & Abou El-Magd, R. (2021). "An Overview of Machine Learning and Deep Learning Techniques for Emotion Recognition". The Journal of Technology Management & Innovation, 16(1), 116-127. doi:10.4067/S0718-27242021000100013
48. Nascimento, C. R., & da Silva, C. L. (2022). "Assessing Emotion Detection Techniques: A Survey and Comparison". Journal of Computer and System Sciences, 79(5), 243-261. doi:10.1007/s00607-021-00831-3.
49. Alam, M., & Ghosh, S. (2018). "Deep Learning for Emotion Recognition: A New Approach". International Journal of Computer Applications, 182(5), 23-28. doi:10.5120/ijca2018916968.
50. Pothin, S., & Gupta, S. (2020). "An Extensive Review on Emotion Recognition Techniques". Journal of Artificial Intelligence Research, 68, 265-306. doi:10.1613/jair.1.12495.
51. Bougouin, A., & Pirotte, A. (2021). "Detection of Emotions in Text: Applications and Theories". Journal of Natural Language Engineering, 27(1), 1-17. doi:10.1017/S1355771820000134.
52. Fang, X., & Zhuang, Z. (2020). "Emotion Recognition in Text Using Machine Learning: An Overview". Soft Computing, 24(24), 19051-19062. doi:10.1007/s00500-020-04761-5.
53. Arora, A., & Sharma, A. (2021). "An Overview of Deep Learning Techniques for Emotion Detection from Text". Universal Journal of Electrical and Electronic Engineering, 9(5), 409-415. doi:10.13189/ijece.2021.090511.
54. Saad, A., & Jabeen, F. (2020). "A Survey on Emotion Detection and Recognition Using Text: Machine Learning Approach". International Journal of Advanced Science and Technology, 29(5), 1683-1690.
55. Maver, M., & Peric, D. (2021). "Emotional Analysis in Text Through Machine Learning Techniques". Journal of MultiCriteria Decision Analysis, 28(5-6), 145-168. doi:10.1002/mcda.2150.
56. Talukder, A., & Das, A. (2021). "Comparative Assessment of Machine Learning Algorithms for Emotion Detection in Text". IET Software, 15(5), 387-399. doi:10.1049/iet-sen.2020.0289.
57. Zhang, L., & Wang, Y. (2020). "Large-Scale Emotion Recognition from Text Using LSTM and Attention". Journal of Ambient Intelligence and Humanized Computing, 11(3), 1095-1106. doi:10.1007/s12652-019-01519-z.
58. Ramya, R. & Kumar, B. (2020). "Emotion Recognition from Text Using Ensemble Learning". Journal of Big Data, 7(1), 1-15. doi:10.1186/s40537-020-00214-5.
59. Arumugam, A., & Sam, M. A. (2022). "Comparative Analysis of Various Techniques for Text Emotion Classification". Journal of Computer Science and Technology, 37(5), 931-949. doi:10.1007/s11390-021-00752-3.
60. Bhatti, M. U., & Zia, S. (2021). "Detecting Emotion in Text: A Comparison of Deep Learning and Traditional Methods". Expert Systems with Applications, 167, 114186. doi:10.1016/j.eswa.2020.114186.
61. Vaquero, J. M., & Castro, M. (2021). "Systematic Review of Emotion Recognition from Text Using Deep Learning". Journal of Computer Languages, Systems & Structures, 64, 101-117. doi:10.1016/j.cl.2021.101117.
62. Raghavendra, M., & Kumar, A. (2021). "Trends in Emotion Recognition from Text: Machine Learning and Neural Approaches". SN Computer Science, 2(4), 1-15. doi:10.1007/s42979-021-00577-8.
63. Ahmed, E., & Rashid, M. (2020). "A Comparative Study of Machine Learning Approaches for Emotion Detection from Text". International Journal of Advanced Computer Science and Applications, 11(5), 1-7. doi:10.14569/IJACSA.2020.0110537.
64. Pant, N. & Kumar, A. (2021). "Sentiment and Emotion Recognition with Machine Learning: A Review and Comparative Study". International Journal of Computational Intelligence Systems, 14(1), 1-12. doi:10.2991/ijcis.d.210315.001
65. Nascimento, C., & Alves, J. (2021). "Machine Learning Approaches for Emotion Detection in Text: A Systematic Review". Artificial Intelligence Review, 54(2), 1207-1245. doi:10.1007/s10462-020-09895-y
66. Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789-5829.
67. Luo, J., Phan, H., & Reiss, J. (2023, January). Fine-tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition. In Proc. Interspeech 2023 (pp. 2413-2417).
68. Li, J., & Xiao, L. (2023, April). Multi-emotion recognition using multi-EmoBERT and emotion analysis in fake news. In Proceedings of the 15th ACM web science conference 2023 (pp. 128-135).
69. GoEmotions: Fine-Grained Emotion Classification, 2022.https://www.kaggle.com/code/enesztrk/goemotions-fine-grained-emotion-classification/input.
70. Souza, M.D., Ananth Prabhu, G. & Kumara, V. Advanced Breast Cancer Detection Using Spatial Attention and Neural Architecture Search (SANAS-Net). SN COMPUT. SCI. 6, 18 (2025). https://doi.org/10.1007/s42979-024-03568-9

71. M. D. Souza, V. Kumara, R. D. Salins, J. J. A Celin, S. Adiga and S. Shedthi, "Advanced Deep Learning Model for Breast Cancer Detection via Thermographic Imaging," 2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Mangalore, India, 2024, pp. 428-433, doi:10.1109/DISCOVER62353.2024.10750727.
72. Luo, Haoran, Tengfei Shao, Shenglei Li, and Tomoji Kishi. "An innovative 3D attention mechanism for multi-label emotion classification: H. Luo et al." Scientific Reports 15, no. 1 (2025): 35951. https://doi.org/10.1038/s41598-025-95804-2.
73. Wu Y, Mi Q, Gao T. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics (Basel). 2025 Jun 27;10(7):418. doi: 10.3390/biomimetics10070418. PMID: 40710231; PMCID: PMC12292624.
74. Deng J, Ren F. Hierarchical Network with Label Embedding for Contextual Emotion Recognition. Research (Wash D C). 2021 Jan 6;2021:3067943. doi: 10.34133/2021/3067943. PMID: 33623915; PMCID: PMC7877378.
75. Ajwa Aslam, Allah Bux Sargano, Zulfiqar Habib, Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks, Applied Soft Computing, Volume 144, 2023, 110494,, ISSN 1568-4946,, https://doi.org/10.1016/j.asoc.2023.110494.
76. Hadi, Muhammad Nur, and Ratri Wulan Sari. "Multi-Label Emotion Detection for Mental Health Monitoring Using Deep CNN and Visual Attention." Journal of Artificial Intelligence and Software Engineering 5, no. 2 (2025): 597-605. http://dx.doi.org/10.30811/jaise.v5i2.6961.
77. Sun, Guoying, Yanan Cheng, Fangzhou Dong, Luhua Wang, Dong Zhao, Zhaoxin Zhang, and Xiaojun Tong. "Multi-label text classification model integrating label attention and historical attention." Knowledge-Based Systems 296 (2024): 111878. https://doi.org/10.1016/j.knosys.2024.111878
78. Liu, Xuan, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. "Emotion classification for short texts: an improved multi-label method." Humanities and Social Sciences Communications 10, no. 1 (2023): 1-9. https://doi.org/10.1057/s41599-023-01816-6.
79. Xue, Jieying, Phuong Nguyen, Minh Nguyen, and Xin Liu. "JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models." In Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), pp. 20-27. 2025.
80. Lian H, Lu C, Li S, Zhao Y, Tang C, Zong Y. A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy (Basel). 2023 Oct 12;25(10):1440. doi: 10.3390/e25101440. PMID: 37895561; PMCID: PMC10606253.