International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

An Enhanced DistilBERT Based Framework for Multi Label Emotion Detection in Social Media Text

Shambhu Nath Saha[1], Rishij Manna[2], Soumya Bhattacharyya[3]
Department of Information Technology, Narula Institute of Technology, Kolkata, India

Abstract. The wide use of social media platforms has resulted in a large volume of user-generated textual content, reflecting different and sometimes overlapping emotions. It is hard to find multiple emotions from such content because social media text has informal writing styles, abbreviations, emojis, hashtags and implicit emotional expressions. Traditional machine learning methods usually rely on hand-crafted features and are limited in capturing complex contextual relationships in text. Recent deep learning models have made significant progress in contextual understanding . However, many existing systems still struggle to identify fine grained emotion specific patterns in multi label emotion classification tasks .

To address these challenges, this work proposes a transformer based framework for multi label emotion detection using the base transformer encoder of DistilBERT architecture. The proposed framework integrates an emotion-specific attention mechanism that allows the model to emphasize emotion-related words across different emotion categories, thus enhancing the representation of contextual features. Furthermore, we present an emoji-aware fine-tuning strategy to improve emotion understanding in emoji-rich social media conversations. Also, a learnable threshold optimization mechanism is also incorporated to dynamically adjust decision boundaries of different emotion classes and improve prediction performance under the imbalanced data condition.

Experimental analysis shows that the proposed framework significantly improves the emotion classification performance compared to the baseline DistilBERT model. The system shows better contextual understanding, better generalization capability and more stable real time emotion prediction over textual and emoji based inputs. We also maintain computational efficiency with the lightweight DistilBERT architecture, making it applicable for real world social media emotion analysis applications.

Keywords: Multi label Emotion Classification, DistilBERT, Emotion Specific Attention, Emoji aware Fine tuning, Social Media Analysis, Natural Language Processing (NLP), Emotion Detection, Contextual Representation Learning.

1 INTRODUCTION

In recent years, social media platforms have become one of the primary channels of digital communication. People use social media sites such as Twitter, Facebook, Instagram, Reddit and online discussion communities to share their opinions, reactions, feelings and day to day experiences on a regular basis. The constant influx of user generated textual data has resulted in a large volume of data containing valuable emotional and behavioral information. Hence the task of emotion detection has attracted a lot of attention in the field of Natural Language Processing (NLP) especially in the fields of sentiment analysis, mental health, recommendation systems, customer feedback analysis, public opinion tracking and human computer interaction etc.

Even though substantial progress has been made in emotion analysis, identifying emotions from text still remains a difficult problem. Human emotions are naturally complex and often depend heavily on context, writing style, and the relationship between words within a sentence. In many real world situations, a single social media post may contain multiple emotions simultaneously. Social media communication further increases the complexity due to the frequent use of slang words, abbreviations, emojis, hashtags, informal grammar, and incomplete sentence structures. These characteristics reduce textual consistency and make emotional interpretation more difficult for computational models.

Earlier emotion detection systems mainly depended on traditional machine learning approaches such as Support Vector Machines (SVM), Naïve Bayes classifiers, and lexicon based methods. Most of these techniques relied heavily on handcrafted features and predefined emotional dictionaries for identifying emotional patterns in text.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Although such approaches produced acceptable performance on limited datasets, they often failed to capture deep semantic meaning and contextual relationships between words. Their effectiveness further decreased when applied to noisy and unstructured social media data.

The development of deep learning models brought major improvements in language understanding tasks. In particular, transformer based architectures such as BERT achieved strong performance across several NLP applications because of their ability to capture contextual dependencies using self attention mechanisms. These models generate context aware representations of words, allowing them to understand sentence meaning more effectively than earlier sequential models. However, full scale BERT models are computationally expensive and require significant memory and processing power, which makes deployment difficult in lightweight or resource constrained environments.

To reduce these limitations, DistilBERT was introduced as a smaller and faster version of BERT while preserving most of its contextual understanding capability. Due to its lightweight architecture and lower computational requirements, DistilBERT has become suitable for large scale text analysis tasks and real time applications. Despite these advantages, many existing emotion detection frameworks still rely mainly on generalized contextual embeddings and fixed classification thresholds. As a result, they may struggle to recognize subtle emotion variations and fine grained emotion specific cues hidden within social media conversations.

Another important aspect of online communication is the growing use of emojis. In many cases, emojis carry strong emotional meaning and sometimes replace textual expressions entirely. They often provide additional context that helps explain the actual emotion behind a sentence. However, several existing approaches do not effectively incorporate emoji information during training, which can negatively affect the quality of emotion prediction, particularly in emoji rich social media text.

To address these challenges, this work proposes a transformer based framework for multi label emotion detection using the DistilBERT architecture. The proposed model integrates an Emotion Specific Attention mechanism that enables the framework to focus on contextually important emotional words for different emotion categories. In addition, an emoji aware fine tuning strategy is introduced to improve the interpretation of implicit emotional signals present in social media conversations. The framework also employs a dynamic threshold optimization mechanism that adaptively determines decision boundaries for different emotion classes, improving prediction performance in multi label classification scenarios.

By combining contextual transformer representations, emotion specific attention learning, emoji aware training, and adaptive threshold optimization, the proposed framework aims to provide a more reliable and computationally efficient approach for multi label emotion detection from social media text.

2 BACKGROUND STUDY

Emotion detection has become an important research area in Natural Language Processing (NLP), especially with the rapid growth of social media platforms and online communication. Earlier studies mainly relied on traditional machine learning methods such as Support Vector Machines (SVM), Naïve Bayes classifiers, and lexicon based approaches for identifying emotions from text [10], [13]. These methods generally depended on handcrafted features and predefined emotional dictionaries. Although they produced acceptable results on smaller and structured datasets, they often struggled to capture deeper contextual meaning and complex emotional relationships in real world social media text.

The introduction of deep learning significantly improved text understanding tasks. Initially, models such as Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks were widely used for sequential text processing because they could learn contextual dependencies better than conventional machine learning approaches [10]. However, these models still faced difficulties in handling long range contextual relationships efficiently.

A major advancement came with the introduction of the Transformer architecture by Vaswani et al. [1]. The Transformer replaced sequential processing with self attention mechanisms, allowing models to capture contextual relationships between words more effectively. This architecture later became the foundation for several modern language models.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Building on the Transformer model, BERT proposed by Devlin et al. [2] introduced bidirectional contextual learning and achieved remarkable performance across various NLP tasks, including sentiment analysis and emotion detection. Since BERT processes text using both left and right context simultaneously, it generates richer semantic representations than earlier language models. However, the original BERT model requires high computational resources and large memory usage, which can make practical deployment difficult.

To reduce these limitations, DistilBERT [3] was introduced as a lightweight version of BERT using knowledge distillation techniques. DistilBERT significantly reduces model size and computational complexity while retaining most of BERT's language understanding capability. Because of its faster inference speed and lower resource requirements, DistilBERT has become suitable for large scale and real time NLP applications.

Several recent studies have explored transformer based architectures for emotion classification tasks. Rezapour [4] showed that transformer models outperform many traditional approaches in contextual emotion understanding. Similarly, Nabiilah [5] compared RoBERTa and DistilBERT for social media emotion classification and reported that lightweight transformer models can provide strong performance while maintaining computational efficiency. RoBERTa [8] further improved transformer based contextual learning through optimized pretraining strategies, while studies by Wolf et al. [6] simplified the practical implementation of transformer architectures for NLP applications.

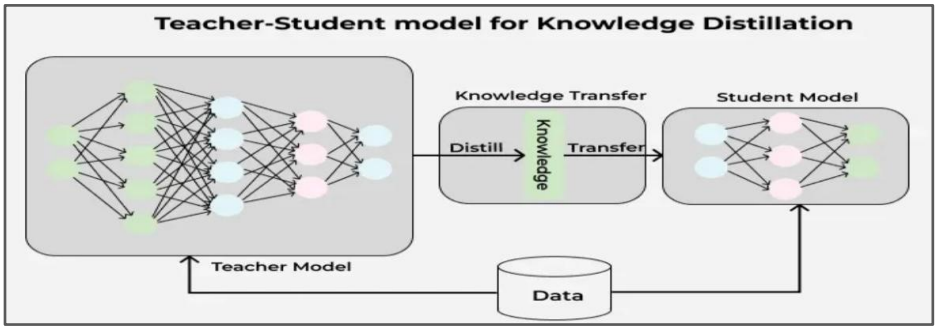

Fig 1. Teacher Student Model Architecture (Knowledge Distillation)

Despite these advancements, several challenges still remain in existing emotion detection systems. Many current approaches rely heavily on generalized sentence embeddings, particularly representations derived from the CLS token. While these embeddings capture overall sentence meaning, they may fail to represent fine grained emotional patterns when multiple emotions appear within the same text. As a result, subtle emotion specific contextual information may not be effectively captured.

Another important limitation is the handling of emojis in social media communication. Emojis often carry strong emotional meaning and sometimes replace textual expressions completely. However, many existing systems mainly focus on textual semantics and do not effectively model emoji emotion relationships, which can reduce the overall quality of emotion prediction.

Class imbalance is another common challenge in multi label emotion classification. Emotional datasets frequently contain uneven distributions where some emotions occur much more often than others. Many earlier studies use fixed probability thresholds during prediction, which may negatively affect the detection of minority emotions and reduce classification consistency.

To address these limitations, the proposed framework integrates a Label Specific Multi Attention mechanism that allows the model to learn separate contextual representations for different emotion categories instead of relying only on a single sentence level embedding. The framework also incorporates emoji aware fine tuning to improve the interpretation of emoji based emotional expressions commonly found in social media text. In addition, an Adaptive Threshold Optimization mechanism is introduced to dynamically determine emotion specific decision boundaries for multi label classification.

Overall, by combining transformer based contextual learning, emotion specific attention modeling, emoji aware training, and adaptive threshold optimization, the proposed framework aims to provide a more effective and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

computationally efficient solution for multi label emotion detection in social media environments

3 METHODOLOGY

The proposed emotion detection framework is developed using a modified DistilBERT architecture combined with an emotion specific attention mechanism and learnable threshold optimization for multi label emotion classification. The model is specially designed for social media text analysis where a single sentence may contain multiple emotions at the same time. Unlike traditional single label classifiers, the proposed system predicts multiple emotional states independently for each input text.

3.1 Dataset Description

1. Training Dataset

The training dataset is based on the GoEmotions dataset and further extended using data collected from Twitter, Facebook, Instagram, YouTube comments, and WhatsApp chats. The final dataset contains unique samples with 29 columns, where one column stores the input text and the remaining 28 columns represent binary emotion labels. Duplicate entries were removed during preprocessing to maintain consistency and improve dataset quality.

2. Testing Dataset

The testing dataset is completely separate from the training dataset and is used only for evaluating model performance on unseen data. It contains unique text samples arranged in the same 29 column format. One column contains the text input, while the remaining 28 columns represent binary emotion labels. The dataset includes samples from multiple social media platforms and different communication styles.

3. Emoji Dataset

The emoji dataset is used during emoji aware fine tuning to help the model understand emoji based emotional expressions. It contains 488 unique emoji samples mapped to the same 28 emotion categories used in the main dataset. Similar to the other datasets, the first column stores the emoji input and the remaining columns contain binary emotion labels for multi label emotion classification.

3.2 Data Preprocessing

Several preprocessing steps were performed to standardize the input data and im prove model performance:

1. The collected dataset was first cleaned and organized before training the model. The text data was converted into a proper format by removing null values and converting all inputs into string format. Since the task is multi label emotion classification, all emotion columns except the text column were selected as target labels.
2. The labels were combined into a numerical array for each sample so that multiple emotions could be assigned to a single sentence. The dataset was then converted into the HuggingFace dataset format for efficient processing and training.
3. For text preprocessing, the DistilBERT tokenizer was used. Each sentence was tokenized into smaller units and converted into numerical token IDs. Padding and truncation were applied with a fixed sequence length to maintain equal input size for the model. Attention masks were also generated to help the model identify meaningful tokens from padded tokens.
4. In the fine tuning stage, emoji based text samples were additionally processed to improve the model performance on social media comments containing emojis. This helped the model learn emotional patterns from both textual and emoji information more effectively.

3.3 Model Architecture

The complete architecture consists of five major stages: text tokenization, contextual feature extraction using DistilBERT, emotion specific attention learning, multi label classification, and adaptive threshold optimization.

Initially, raw social media comments are provided as input to the tokenizer. The DistilBERT tokenizer converts each sentence into smaller subword tokens using WordPiece tokenization. This helps the model handle unknown words, spelling variations, emojis, abbreviations, and informal internet language more effectively.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

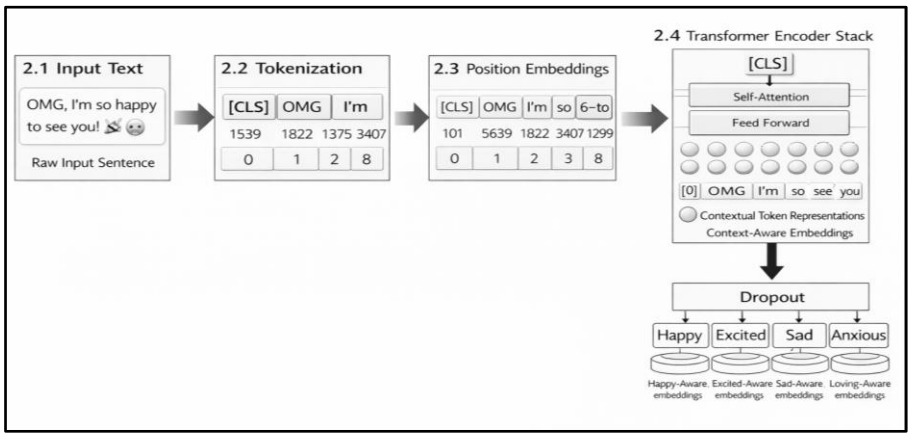

Fig 2. Transforming Raw Input Text into Contextualized Representation

Each token is mapped into a numerical token ID because deep learning models cannot process raw textual data directly.

Special tokens are also added during tokenization. The [CLS] token is added at the beginning of the sequence to represent sentence level information, while the [SEP] token is used to indicate the end of the sentence. Padding and truncation are applied to maintain equal sequence length for all samples during batch processing.

For every input sentence, two important vectors are generated:

1. Input IDs
2. Attention Masks

The attention mask helps the model identify valid tokens and ignore padded positions during computation.

After preprocessing, the tokenized inputs are passed into the pretrained DistilBERT encoder. DistilBERT is a compressed version of the BERT architecture that retains most of the language understanding capability while reducing computational complexity. It uses Transformer encoder layers and self attention mechanisms to capture contextual information from text.

The hidden representation generated by DistilBERT can be mathematically represented as:

$$H \in R^{(B \times S \times D)}$$

where:

B = batch size
S = sequence length
D = hidden embedding dimension

Each token embedding generated by DistilBERT contains contextual semantic information. Unlike traditional word embeddings where each word has a fixed meaning, contextual embeddings change depending on surrounding words. This helps the model understand emotional meaning more accurately in different contexts.

The Transformer encoder inside DistilBERT mainly relies on the self attention mechanism. Self attention allows every token to interact with all other tokens present in the sentence. The attention calculation is based on Query (Q), Key (K), and Value (V) matrices.

The **self attention operation** is mathematically defined as:

$$Attention(Q,K,V) = Softmax((QK^T) / \sqrt{d_k})V$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

where d_k represents the dimension of the key vectors.

	I	made	a	sweet	indian	rice	dish	called
I	–	42.0%	18.0%	6.0%	4.0%	12.0%	10.0%	8.0%
made	30.0%	–	20.0%	10.0%	6.0%	14.0%	12.0%	8.0%
a	15.0%	18.0%	–	22.0%	10.0%	18.0%	12.0%	5.0%
sweet	4.0%	6.0%	10.0%	–	28.0%	30.0%	18.0%	4.0%
indian	3.0%	5.0%	6.0%	26.0%	–	35.0%	22.0%	3.0%
rice	5.0%	10.0%	2.5%	25.0%	20.0%	–	30.0%	7.5%
dish	0.7%	1.1%	1.4%	36.3%	11.0%	19.3%	–	30.2%
called	2.0%	8.0%	1.5%	6.0%	4.0%	10.0%	48.0%	–

Fig 3. Contextual Representation

This mechanism helps the model learn long range dependencies between words. For example, the emotional meaning of a sentence often depends on relationships between distant words, negations, or contextual expressions. Self attention helps preserve these dependencies more effectively than traditional recurrent neural networks.

Although DistilBERT provides powerful contextual representations, the standard CLS token representation does not always capture emotion specific features efficiently for multi label emotion detection. To solve this limitation, an Emotion Attention layer is introduced in the proposed architecture.

Instead of assigning equal importance to all tokens, the proposed attention mechanism learns separate attention vectors for each emotion category. This allows the model to focus on different emotional words for different emotions.

The trainable emotion attention matrix is represented as:

$$A \in R^{(E \times D)}$$

where:

E = number of emotion classes
D = hidden dimension

The interaction between hidden token representations and emotion specific attention vectors is computed using tensor multiplication.

The attention score calculation is defined as:

$$Score = H \times A^T$$

This operation produces attention scores for every token with respect to every emotion category. The Softmax function is then applied across the sequence dimension to normalize the scores into probability distributions.

$$\alpha_i = e^{(s_i)} / \Sigma_j e^{(s_j)}$$

where α_i represents the normalized attention weight.

These attention weights indicate how important a particular token is for detecting a specific emotion. Tokens carrying stronger emotional information receive higher attention values, while less important words receive lower

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

weights.

The weighted emotion representation is computed by combining hidden states using the learned attention weights. This operation helps the model generate separate semantic representations for each emotion category.

Compared to the traditional CLS token approach, this mechanism provides better emotion localization because the model explicitly learns where emotional information is present inside the sentence.

After attention computation, a dropout layer is applied to reduce overfitting during training. Dropout randomly deactivates neurons during forward propagation so that the model does not become overly dependent on specific features.

The dropout operation can be represented as:

$$h' = h \cdot m$$

where m represents a randomly generated binary mask.

The final emotion representations are then passed through a fully connected linear classification layer. A separate logit value is generated for every emotion category.

$$z = Wx + b$$

where:

W = learnable weight matrix
x = input feature vector
b = bias term

Since the task is multi label classification, the Sigmoid activation function is applied independently to each output neuron instead of using Softmax. Softmax is generally used for mutually exclusive classes, whereas Sigmoid allows multiple emotions to be predicted simultaneously.

The Sigmoid activation function is defined as:

$$P(y_i) = \sigma(z_i) = 1 / (1 + e^{-z_i})$$

The output probability for each emotion lies between 0 and 1. Higher probability values indicate stronger confidence for the presence of a particular emotion.

For model optimization, Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) is used. This loss function is suitable for multi label classification because it independently calculates loss for each emotion category.

The BCE loss function is mathematically represented as:

$$L = (1/N) \sum_i [\, y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \,]$$

where:

y_i = actual label
p_i = predicted probability

The AdamW optimizer is used for parameter optimization during training. AdamW combines adaptive learning rate optimization with weight decay regularization, which helps improve convergence and prevents overfitting.

The parameter update process is given as:

$$\theta_t = \vartheta_{t-1} - \eta \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$$

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

where:

η = learning rate
$\hat{m}_t$ = bias corrected first moment estimate
$\hat{v}_t$ = bias corrected second moment estimate

To further improve prediction quality, a learnable threshold optimization stage is introduced after fine tuning. In most multi label classification systems, a fixed threshold value of 0.5 is used for all classes. However, emotion distributions are highly imbalanced, and different emotions require different confidence levels for accurate prediction.

To solve this issue, separate trainable thresholds are learned for each emotion category.

The threshold vector is represented as:

$$T \in R^E$$

where E denotes the number of emotion classes.

During threshold learning, all pretrained model parameters are frozen and only threshold values are optimized. This reduces computational cost and prevents disturbance to previously learned semantic representations.

A differentiable thresholding function is used so that gradients can still propagate during optimization.

The differentiable threshold operation is defined as:

$$\hat{y} = \sigma((p \quad t) \times 10)$$

where:

p = predicted probability
t = learnable threshold value

The multiplication factor sharpens the sigmoid transition and improves threshold sensitivity during optimization.

The threshold optimization stage helps the model adapt better to class imbalance, low frequency emotions, and varying confidence distributions across different emotion categories.

In the final stage, emoji enhanced fine tuning is performed using emoji rich social media comments. Emojis often carry strong emotional meaning that may not be fully captured through plain text alone. Therefore, additional emoji based samples are introduced during fine tuning to improve emotional representation learning.

The emoji preprocessing stage extracts emojis using regular expression based pattern matching and incorporates them into the training process. This helps the model learn emotion patterns from both textual semantics and symbolic emotional expressions.

Overall, the proposed architecture combines contextual language understanding, emotion focused attention learning, adaptive threshold optimization, and emoji aware fine tuning to improve multi label emotion detection performance on real world social media data.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

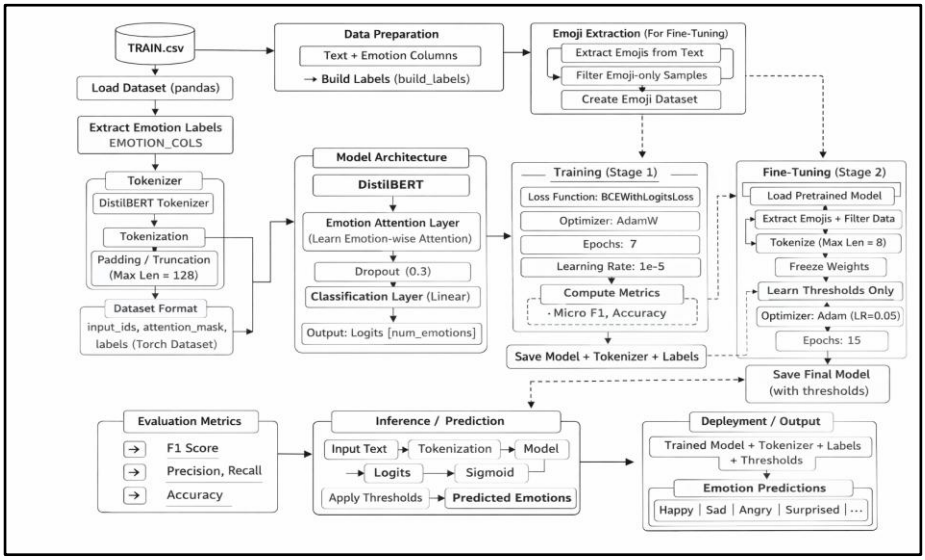

Fig 4. Proposed System Architecture

4 RESULTS AND DISCUSSIONS

The performance of the proposed Emotion Specific Attention Model was evaluated on three different datasets: Training Dataset, Testing Dataset, and Emoji Dataset. The obtained results were compared with the baseline DistilBERT model to analyze the effectiveness of the proposed modifications.

The overall performance comparison of the proposed model across different datasets is presented in Fig. 5.

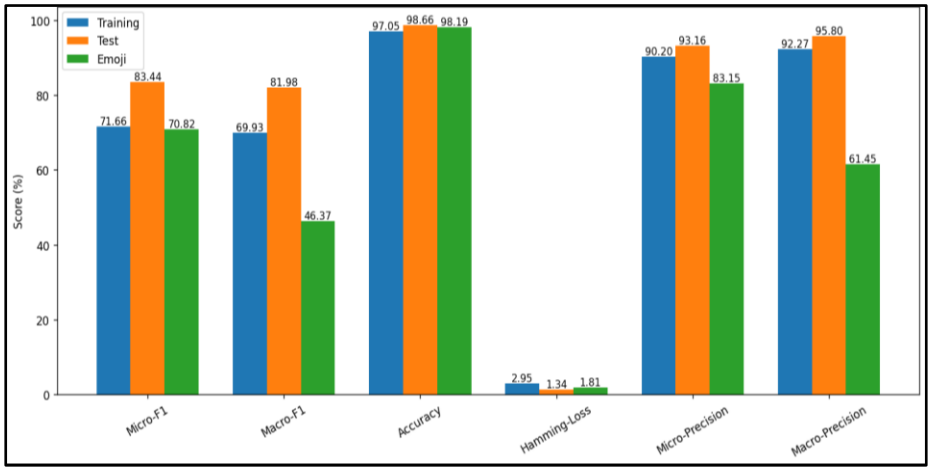

Fig 5: Emotion Specific Attention model performance across datasets

The model achieved strong classification performance on both training and testing datasets with noticeable improvement over the emoji dataset as well. The Micro F1 score reached 83.44% on the testing dataset, while Macro F1 score reached 81.98%, indicating balanced performance across multiple emotion classes. The model also achieved high accuracy values above 96% across all datasets. In addition, lower Hamming Loss values indicate that the proposed model produced fewer incorrect label predictions during multi label classification.

The improvement in performance mainly comes from the proposed Emotion Attention mechanism. Unlike the traditional CLS token based representation, the attention layer learns separate contextual importance for each emotion category. This helps the model focus on emotionally relevant words more effectively, improving multi label emotion understanding.

The baseline DistilBERT model results are shown in Fig. 6.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

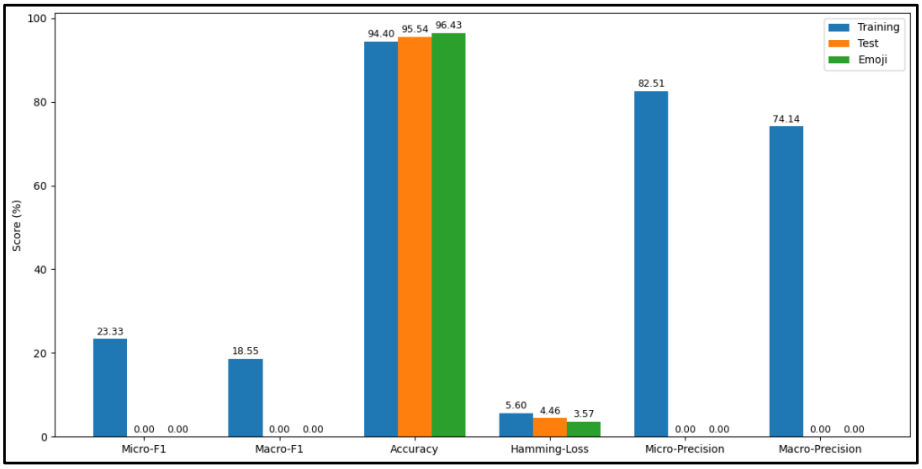

Fig 6: Baseline model performance across datasets

From the results, it can be observed that the baseline model struggled significantly on the testing and emoji datasets. Although the training accuracy remained high, the Micro F1 and Macro F1 scores on testing and emoji datasets dropped close to zero. This indicates poor generalization capability and inability to properly capture emotion specific contextual information.

A major limitation of the baseline model is that it relies mainly on a single CLS token representation for classification. This representation often fails to capture fine grained emotional dependencies present in social media text. As a result, the baseline model produced unstable predictions for unseen emotional patterns and emoji rich inputs.

A detailed per emotion performance comparison between the baseline model and the proposed Emotion Specific Attention Model is presented in Fig. 7.

Base Model

Emotion	Training Dataset	Test Dataset	Emoji Dataset
admiration	21.88	0.0	0.0
amusement	11.48	0.0	0.0
anger	9.93	0.0	0.0
annoyance	16.31	0.0	0.0
approval	61.73	0.0	0.0
caring	14.99	0.0	0.0
confusion	14.55	0.0	0.0
curiosity	8.94	0.0	0.0
desire	14.88	0.0	0.0
disappointment	16.04	0.0	0.0
disapproval	7.36	0.0	0.0
disgust	2.71	0.0	0.0
embarrassment	15.45	0.0	0.0
excitement	10.87	0.0	0.0
fear	7.91	0.0	0.0
gratitude	25.41	0.0	0.0
grief	11.76	0.0	0.0
joy	26.98	0.0	0.0
love	15.15	0.0	0.0
nervousness	20.80	0.0	0.0
neutral	24.39	0.0	0.0
optimism	61.74	0.0	0.0
realization	10.81	0.0	0.0
pride	0.18	0.0	0.0
relief	58.40	0.0	0.0
remorss	12.21	0.0	0.0
sadness	11.74	0.0	0.0
surprise	4.81	0.0	0.0

Emotion-Specific Attention Model

Emotion	Training Dataset	Test Dataset	Emoji Dataset
admiration	72.16	89.36	28.57
amusement	78.34	91.35	57.14
anger	83.72	100.00	100.00
annoyance	65.86	86.49	40.00
approval	86.11	86.03	65.00
caring	78.36	98.54	33.33
confusion	59.67	72.57	57.14
curiosity	71.26	80.40	100.00
desire	82.08	100.00	33.33
disappointment	41.82	48.13	40.00
disapproval	77.25	89.41	96.55
disgust	78.79	91.15	57.14
embarrassment	63.66	80.33	33.33
excitement	82.35	100.00	28.57
fear	75.60	91.38	85.71
gratitude	79.82	99.99	0.00
grief	58.44	69.21	66.67
joy	76.13	91.37	31.58
love	68.46	85.67	46.15
nervousness	53.05	72.26	66.67
neutral	64.80	77.92	82.10
optimism	86.28	86.08	0.00
realization	39.46	41.11	0.00
pride	83.38	99.43	94.12
relief	82.32	77.48	40.00
remorse	75.32	96.28	80.00
sadness	55.54	61.46	57.14
surprise	69.54	77.62	60.00

Fig 7: Per Emotion F1 comparison between Baseline Model and Emotion Specific Attention Model

The results clearly show that the proposed model achieved substantial improvement across most emotion categories. Emotions such as anger, amusement, curiosity, excitement, pride, and fear showed major performance gains compared to the baseline model.

The proposed model achieved very high F1 scores on several emotions in the testing dataset, including anger

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

(100%), excitement (100%), desire (100%), gratitude (99.99%), and pride (99.43%). This demonstrates the effectiveness of the emotion specific attention mechanism in learning discriminative emotional representations.

However, certain emotions such as grief, sadness, realization, and disappointment achieved relatively lower performance compared to dominant emotions. This may be caused by class imbalance, semantic overlap between emotions, and limited emotional context available in some samples. Despite these challenges, the proposed model still performed considerably better than the baseline architecture.

The real time inference performance of the proposed Emotion Specific Attention Model is shown in Fig. 8.

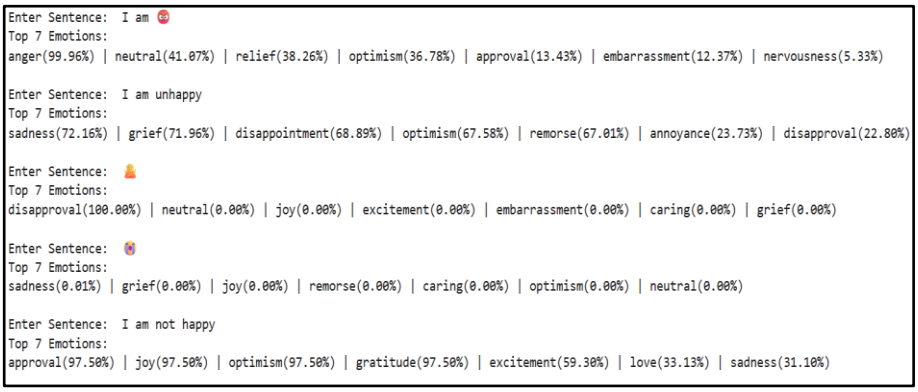

Fig 8 : Real time performance of Emotion Specific Attention Model

The model successfully detected multiple emotions from different types of textual and emoji based inputs. For example, positive inputs generated high confidence scores for emotions such as joy, optimism, gratitude, and approval, while negative inputs produced stronger predictions for sadness, grief, disappointment, and remorse. The model also demonstrated the ability to process emoji only inputs, indicating improved emotional understanding of symbolic expressions.

In contrast, the baseline model inference outputs shown in Fig. 9 produced unstable and inconsistent predictions.

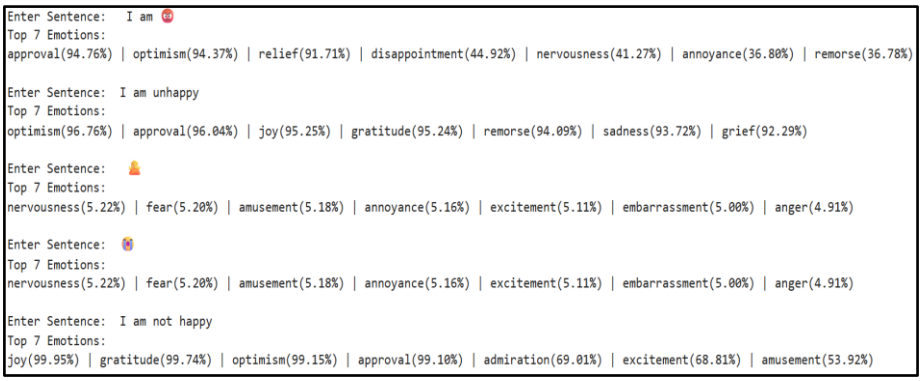

Fig 9 : Real time performance of Baseline Model

Several outputs contained unrelated emotions with unrealistic confidence scores, especially for emoji inputs and negative emotional statements. This further confirms that the baseline model lacks sufficient emotion localization capability and contextual emotional understanding.

Another important contribution of the proposed system is the learnable threshold optimization mechanism. Instead of using a fixed threshold for all emotion classes, separate trainable thresholds were learned for individual emotions. This helped improve classification balance and reduced incorrect predictions caused by varying emotion distributions. The threshold optimization stage also improved model adaptability for minority emotion classes.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The emoji enhanced fine tuning stage further improved the robustness of the system on real world social media comments. Since social media users frequently express emotions through emojis, the additional emoji based training helped the model capture emotional semantics beyond plain textual information.

Overall, the obtained experimental results demonstrate that the proposed Emotion Specific Attention Model significantly outperformed the baseline DistilBERT model in terms of classification accuracy, contextual understanding, generalization capability, and real time emotion prediction performance.

5 CONCLUSIONS

In this research, an enhanced DistilBERT based multi label emotion detection system was proposed for analyzing social media comments. The proposed architecture combined contextual language understanding with an emotion specific attention mechanism to improve the detection of multiple emotions from a single text input. In addition, emoji enhanced fine tuning and learnable threshold optimization were introduced to improve the model performance on real world social media data.

The experimental results demonstrated that the proposed model significantly outperformed the baseline DistilBERT model across training, testing, and emoji datasets. The emotion specific attention mechanism helped the model focus on emotionally important words more effectively, resulting in better contextual understanding and improved classification accuracy. The threshold learning approach also improved prediction balance by adapting separate confidence thresholds for different emotion categories.

The model showed strong performance in both quantitative evaluation metrics and real time inference testing. It was able to identify multiple emotions simultaneously from text and emoji based inputs with higher stability and better generalization capability compared to the baseline model.

Although the proposed system achieved promising results, certain emotions with limited training samples still showed relatively lower performance due to class imbalance and semantic similarity between emotions. In future work, larger multilingual datasets, advanced attention mechanisms, and multimodal emotional features such as audio and visual signals can be incorporated to further improve emotion understanding and robustness.

Overall, the proposed Emotion Specific Attention Model provides an effective and practical approach for multi label emotion detection in modern social media applications.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention Is All You Need,” Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.
[2] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, “BERT: Pre training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.
[3] V. Sanh, L. Debut, J. Chaumond and T. Wolf, “DistilBERT: A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter,” Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC²), Vancouver, Canada, 2019.
[4] M. Rezapour, “Emotion Detection with Transformers: A Comparative Study,” 2024.
[5] G. Z. Nabiilah, “Effectiveness Analysis of RoBERTa and DistilBERT in Emotion Classification Task on Social Media Text Data,” Journal EMACS (Engineering, Mathematics and Computer Science), vol. 7, no. 1, pp. 45–50, 2025.
[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, et al., “Transformers: State of the Art Natural Language Processing,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, 2020, pp. 38–45.
[7] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, “Improving Language Understanding by Generative Pre Training,” OpenAI Technical Report, 2018.
[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692, 2019.
[9] J. Howard and S. Ruder, “Universal Language Model Fine tuning for Text Classification,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018, pp. 328–339.
[10] S. Hochreiter and J. Schmidhuber, “Long Short Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] A. Severyn and A. Moschitti, “Twitter Sentiment Analysis with Deep Convolutional Neural Networks,” Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 2015, pp. 959–962.
[12] F. Barbieri, J. Camacho Collados, L. Espinosa Anke and M. Neves, “TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification,” Findings of EMNLP, 2020, pp. 1644–1650.
[13] E. Cambria, D. Das, S. Bandyopadhyay and A. Feraco, “A Practical Guide to Sentiment Analysis,” Socio Affective Computing Series, Springer, 2017.
[14] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria and R. Zimmermann, “ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Belgium, 2018, pp. 2594–2604.

[15] S. Poria, E. Cambria, D. Hazarika and N. Majumder, “Context Dependent Sentiment Analysis in User Generated Videos,” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, Canada, 2017, pp. 873–883.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)