scientific reports

Check for updates

OPEN

Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition

Fatimah Alqarni[1,2], Alaa Sagheer[1,2]✉, Amira Alabbad[1] & Hala Hamdoun[1]

Emotion recognition in text is a fundamental task in natural language processing, underpinning applications such as sentiment analysis, mental health monitoring, and content moderation. Although transformer-based models like RoBERTa have advanced contextual understanding in text, they still face limitations in identifying subtle emotional cues, handling class imbalances, and processing noisy or informal input. To address these challenges, this paper introduces Emotion-Aware RoBERTa, an enhanced framework that integrates an Emotion-Specific Attention (ESA) layer and a TF-IDF based gating mechanism. These additions are designed to dynamically prioritize emotionally salient tokens while suppressing irrelevant content, thereby improving both classification accuracy and robustness. The model achieved 96.77% accuracy and a weighted F1-score of 0.97 on the primary dataset, outperforming baseline RoBERTa and other benchmark models such as DistilBERT and ALBERT with a relative improvement ranging from 9.68% to 10.87%. Its generalization capability was confirmed across two external datasets, achieving 88.03% on a large-scale corpus and 65.67% on a smaller, noisier dataset. An ablation study revealed the complementary impact of the ESA and TF-IDF components, balancing performance and inference efficiency. Attention heatmaps were used to visualize ESA's ability to focus on key emotional expressions, while inference-time optimizations using FP16 and Automatic Mixed Precision (AMP) reduced memory consumption and latency. Additionally, McNemar's statistical test confirmed the significance of the improvements over the baseline. These findings demonstrate that Emotion-Aware RoBERTa offers a scalable, interpretable, and deployment-friendly solution for fine-grained emotion recognition, making it well-suited for real-world NLP applications in emotion-aware systems.

Keywords Natural language processing, Transformers, Deep learning, Emotion recognition, Attention mechanism, Text classification, Attention heatmaps, FP16 and AMP, McNemar test

The rapid growth of digital communication has resulted in an unprecedented volume of textual data, creating significant opportunities for analyzing emotions and individuals' opinions. Social media platforms, news articles, and other online communication channels contain emotional signals that provide valuable insights for applications such as customer sentiment analysis, mental health monitoring, and the recognition of harmful content, including cyberbullying and hate speech[1]. However, accurately recognizing emotions in text remains a complex challenge due to linguistic ambiguity, context dependencies, and cultural nuances[2].

Recent advancements in Natural Language Processing (NLP), particularly with transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers)[3] and RoBERTa (Robustly Optimized BERT Pretraining Approach)[4], have significantly enhanced emotion recognition and sentiment analysis by effectively capturing contextual relationships between words and phrases[5]. Despite these advancements, fine-grained emotion recognition remains a challenging task, particularly in real-world, informal text domains where emotions are often expressed through implicit cues, figurative language, and context-dependent phrasing[6]. Traditional RoBERTa, while robust in sentiment classification, still struggles with understanding sarcasm, irony, and indirect emotional expressions due to the intricate linguistic structures and subtle cues inherent in these phenomena[7].

[1]Department of Computer Science, College of Computer Sciences and Information Technology, King Faisal University, Hofuf, Saudi Arabia. [2]These authors contributed equally: Fatimah Alqarni and Alaa Sagheer. ✉email: asagheer@kfu.edu.sa

For example, in the sentence, "*Oh great, another meeting at 8 AM. I just LOVE it*" RoBERTa may incorrectly classify "love" as a positive sentiment, failing to recognize the sarcasm intended. This limitation stems from the model's reliance on lexical features without deeper contextual reasoning, which is necessary for interpreting nuanced emotional cues[8]. Existing studies predominantly focus on broad sentiment classification (positive, negative, or neutral), often neglecting the complexities of emotions such as frustration, sarcasm, or mixed affective states. These nuanced expressions are particularly prevalent in social media and conversational text, where informal language, abbreviations, and contextual ambiguity further complicate accurate classification[9,10].

Another critical challenge is the generalization capability of transformer-based models in emotion recognition. Fine-tuned models often exhibit strong performance on the datasets they are trained on but struggle to maintain accuracy when applied to unseen datasets with different linguistic distributions[11]. Additionally, handling noisy and domain-specific text, including informal language, slang, and context-dependent terminology, poses a significant obstacle to extracting meaningful emotional representations[12]. Furthermore, class imbalance remains a persistent issue, where rare emotions such as surprise and fear are underrepresented in training datasets, leading to biased predictions and poor recall for these categories[13].

Despite the progress in transformer-based architectures, current models lack targeted mechanisms to address these limitations effectively. There is a clear need for an improved approach that enhances RoBERTa's capability to handle nuanced emotions, improve generalization across datasets, and mitigate class imbalance, all while maintaining computational efficiency. To bridge this gap, this paper introduces an enhanced version of RoBERTa, called Emotion-Aware RoBERTa, designed to refine emotion recognition in complex and linguistically diverse contexts.

The proposed enhancement integrates advanced preprocessing techniques, including slang-aware tokenization and tailored normalization strategies to better handle noisy text. To address class imbalance, a random oversampling mechanism is applied to increase the representation of underrepresented emotions, ensuring a more balanced training process. A key innovation of this approach is the introduction of an Emotion-Specific Attention (ESA) layer, which prioritizes emotionally salient tokens by dynamically adjusting attention weights, enhancing the model's ability to distinguish subtle emotional variations while preserving syntactic and semantic coherence. Additionally, the TF-IDF based gating mechanism refines attention distribution by filtering out low-information tokens while amplifying high-relevance emotional cues, further improving classification accuracy.

Beyond these enhancements, the proposed modifications significantly improve the model's generalization capability, ensuring robust performance across multiple datasets with different linguistic characteristics. This enables the proposed model to better adapt to real-world applications such as sentiment analysis, mental health monitoring, and human-computer interaction, addressing key challenges in emotion recognition with a scalable and domain-adaptive solution.

To ensure a fair evaluation, the empirical experiments in this study demonstrate that the proposed model achieved an accuracy of 96.77%, outperforming the baseline RoBERTa by 10.87% under unified experimental settings, including the same dataset. Additionally, the proposed model surpassed other transformer-based counterparts, such as ALBERT[14] and DistilBERT[15], which achieved accuracy as 85.85% and 87.10%, respectively. Furthermore, the model's generalization capability was assessed on two external datasets of varying sizes and linguistic characteristics, where it exhibited strong and consistent performance across different domains, class distributions, and communication styles. These results underscore the robustness of Emotion-Aware RoBERTa in adapting to diverse datasets while maintaining high classification accuracy.

By leveraging RoBERTa's transformer-based architecture and integrating ESA and TF-IDF based gating mechanisms, this paper addresses key limitations in emotion recognition, particularly in handling class imbalance, subtle emotional variations, and noisy text. The proposed enhancements advance the development of more effective emotion-aware NLP systems, improving their scalability and applicability to real-world sentiment analysis, mental health monitoring, and content moderation tasks[16].

Accordingly, the contributions of this paper can be summarized as follows:

- Proposing an emotion-specific attention (ESA) layer that enhances RoBERTa's ability to capture fine-grained emotional cues. The ESA layer prioritizes emotionally salient tokens, thereby improving the detection of nuanced and rare emotions through adaptive attention reweighting.
- Integrating a TF-IDF based gating mechanism to filter out low-information tokens and enhance the emotional relevance of input features. This gating reduces noise, sharpens attention focus, and improves classification robustness, especially in the presence of informal or unstructured text.
- Designing a domain-aware preprocessing pipeline that includes slang-aware tokenization, text normalization, and random oversampling. This pipeline improves model resilience to noisy, imbalanced, and informal language, which is common in real-world emotion-rich datasets.
- Conducting a comprehensive multi-dataset evaluation to validate the model's generalization across three benchmark datasets with diverse linguistic styles and domain characteristics. The proposed model consistently outperforms strong transformer-based baselines in emotion classification tasks.
- Demonstrating the Emotion-Aware RoBERTa as a scalable, interpretable, and deployment-ready solution for real-world applications in emotion-aware NLP, including sentiment analysis, mental health monitoring, content moderation, and human-computer interaction.

The remainder of this paper is organized as follows. "Literature review" section reviews recent literature on emotion recognition and transformer-based models. "Problem statement and motivation" section presents the problem statement and the motivation behind this paper. "Methodology" section details the proposed methodology, including the integration of the ESA layer and TF-IDF gating mechanism. "Datasets and experimental

setup" section outlines the datasets used, preprocessing steps, and experimental setup. "Experimental results and discussion" section reports and discusses the experimental results, featuring performance comparisons, ablation studies, benchmarking against transformer models, and inference-time optimization using FP16 and AMP. "Limitations and future directions" section outlines limitations and proposes future research directions. "Conclusion" section concludes the paper with a summary of key findings and implications for emotion-aware NLP applications.

Literature review

The field of sentiment analysis and emotion recognition has gained significant attention in recent years, driven by advancements in NLP and deep learning techniques. Traditional approaches, such as lexicon-based methods and classical machine learning algorithms, have shown limited success due to their inability to capture contextual dependencies and nuanced emotional expressions. With the rise of transformer-based models, particularly BERT[3] and RoBERTa[4], emotion recognition has seen remarkable improvements, benefiting from their ability to process long-range dependencies, enhance contextual understanding, and encode rich semantic representations.

Despite these advancements, several challenges remain unresolved. Transformer-based models often struggle with class imbalance, where rare emotions such as *surprise* and *fear* are underrepresented in training datasets, leading to biased predictions[17]. Additionally, handling informal and noisy text, including code-mixed language and social media slang, poses significant difficulties[18]. Many existing studies primarily focus on broad sentiment classification (positive, negative, or neutral) while overlooking more complex emotions such as sarcasm, frustration, and mixed sentiments, which are prevalent in real-world applications[9,10]. Recent research has explored various techniques for enhancement, including model fine-tuning, and data augmentation strategies, to improve the generalization and robustness of transformer-based models. This section provides a review of the recent approaches, categorizing them into two key areas: (1) transformer-based models for general text classification and (2) advancements in emotion recognition.

General text recognition

General text recognition tasks focus on classifying textual content into specific categories, such as fake news, hate speech, or general intent recognition. Several studies have explored improvements in these areas. Kitanovski et al.[19] compared the performance of DistilBERT and RoBERTa for fake news detection across two datasets. Their findings showed that RoBERTa outperformed DistilBERT, particularly in recall for the fake news class, making it more effective in this domain. However, a major limitation was the models poor generalization to new datasets, which reduced their effectiveness beyond training data.

Chen and Ye[20] investigated satire detection using a BERT+LSTM architecture. Their study focused on the challenges of recognizing sarcasm and satire in news headlines, where contextual understanding is critical. While the model performed well on shorter, straightforward headlines, it struggled with longer, more complex sentences where satirical cues were more subtle. Their study highlighted the need for models capable of managing long-range dependencies and nuanced language. Choudhry et al.[21] adopted a multitask approach that incorporated emotion recognition into fake news and rumor detection tasks. Their emotion-aware framework improved the representation of emotional contexts surrounding misinformation, enhancing model performance. However, the subjective nature of emotion labels introduced inconsistencies across datasets, complicating the model's training process.

Lai and Zhang[22] proposed a hybrid model integrating RoBERTa and TextCNN for classifying government affairs messages. The combination of RoBERTa's contextual embeddings with TextCNN's feature extraction led to substantial improvements in classification accuracy. Despite this, the unstructured nature of government messaging and noisy data necessitated extensive preprocessing to mitigate redundancy and non-semantic information. Santhiya et al.[23] explored hate speech and offensive language detection on social media using BERT-based models and GloVe embeddings. The integration of CNN further boosted performance. However, class imbalance in datasets posed significant challenges, with oversampling techniques introducing noise and increasing model complexity. Addressing this imbalance remains a crucial area for future research.

Mladenovic et al.[24] examined the formidable and complex challenge of insider threats to organizational security. The research involves six experiments using email, HTTP, and file content data. The experiments' focus was on recognizing the sentiment and context of malicious actions, which are considered less prone to change compared to commonly tracked metrics like location and time of access. The novelty here is in using metaheuristic-optimized machine learning classifiers to combat insider threats.

Advancements in emotion recognition

Emotion recognition focuses on identifying the emotional tone or specific emotions, such as joy, anger, or sadness, conveyed in text. Transformer-based models, particularly BERT and RoBERTa, have demonstrated significant advancements in this task due to their ability to capture contextual dependencies and long-range relationships between words. Wang et al.[25] applied RoBERTa for sentiment classification in Chinese, leveraging data augmentation techniques to mitigate the limitations of training data. Their approach yielded notable accuracy improvements, especially in binary sentiment classification. However, the noise introduced by data augmentation methods negatively impacted performance consistency, highlighting the need for more sophisticated techniques to enhance recognition.

Maity et al.[17] developed a multitasking framework for detecting cyberbullying in Hindi-English (Hinglish) tweets, incorporating sentiment and emotion analysis as auxiliary tasks. By leveraging BERT and emoji data, their model surpassed single-task models, demonstrating the interaction between sentiment, emotion, and cyberbullying detection. On the other hand, code-mixed languages complicated data processing and model interpretation. Saha et al.[26] introduced a multitask model emphasizing speech act classification in tweets.

Integrating sentiment and emotion detection improved communicative intent identification. However, the unstructured and noisy nature of tweets, including emoticons and non-standard language, hindered consistent classification, underscoring the need for refined preprocessing methods.

Bashir et al.[18] addressed emotion recognition in low-resource languages, applying deep neural networks to analyze emotions in Urdu. They introduced the Urdu Nastalique Emotions Dataset (UNED), annotated with six emotions. While their model achieved an F1 score of 85% on sentence-level data, its performance on paragraph-level data was notably lower 50%, reflecting the difficulty of capturing emotions in longer text. Despite these sincere efforts, significant challenges persist in accurately capturing complex emotional nuances such as sarcasm, mixed sentiments, and subtle affective expressions. Additional difficulties arise from class imbalance and domain-specific variations, particularly in noisy and informal text. These limitations highlight the continued need for advancing model architectures and preprocessing strategies to enhance the accuracy, robustness, and generalizability of emotion recognition systems.

Problem statement and motivation

The task addressed in this paper is emotion recognition in text, a critical component in a wide range of real-world applications, including sentiment analysis, mental health monitoring, customer support automation, human-computer interaction, and adaptive education systems. Emotion recognition becomes particularly challenging in informal and diverse forms of communication, such as conversational language, social media posts, and short reviews. These forms often contain subtle and context-dependent emotional cues, including sarcasm, frustration, or ambiguous expressions of joy and sadness, which are difficult to detect using conventional sentiment analysis techniques[27].

Although transformer-based models such as RoBERTa have significantly improved text classification by capturing contextual relationships and long-range dependencies, several limitations remain that hinder their deployment in real-world settings. One persistent issue is limited generalization: models fine-tuned on specific datasets often fail to maintain high performance when evaluated on unseen or cross-domain data due to overfitting[11,28]. Additionally, noisy, domain-specific, or informal text, characterized by slang, abbreviations, or inconsistent syntax, complicates the extraction of meaningful representations, especially when models are trained on more standardized corpora[12,29].

Another critical obstacle is class imbalance. In emotion-labeled datasets, some emotions, such as surprise or fear, are underrepresented. Standard deep learning models often exhibit bias toward majority classes, resulting in degraded recall for minority emotions and skewed overall performance[13]. Furthermore, even when classification accuracy is high, misclassification remains prevalent between semantically similar emotions (e.g., joy vs. love), and these errors are rarely revealed by aggregate metrics alone. These limitations hinder the applicability of current models in diverse linguistic and contextual settings, making emotion recognition in complex real-world scenarios an ongoing research challenge. This paper addresses these challenges by integrating two key components: an Emotion-Specific Attention layer to dynamically emphasize emotionally relevant tokens, and a TF-IDF based gating mechanism to suppress noise and improve inference efficiency. This novel architecture is described in detail in the following sections.

Methodology

This paper proposes Emotion-Aware RoBERTa, an enhanced emotion recognition framework that integrates an ESA layer and a TF-IDF based gating mechanism to improve the model's ability to capture nuanced emotional expressions. The proposed approach enhances RoBERTa for emotion recognition by incorporating an emotion-specific attention layer and refining data preprocessing through TF-IDF based gating. The methodology consists of several key stages, beginning with text preprocessing. Following this, tokenization and data preparation are carried out using Hugging Face's Roberta TokenizerFast, ensuring dynamic padding and optimal tokenization strategies. A data loader with padding collators is employed to maximize computational efficiency by dynamically adjusting batch sizes based on available GPU memory.

To address the issue of irrelevant or low-impact words affecting the classification process, a TF-IDF based gating mechanism is applied. The processed input is then passed through the ESA layer, which enhances the model's ability to capture subtle emotional cues by dynamically adjusting attention weights for emotion-bearing words. This mechanism strengthens the differentiation of closely related emotions while improving recall for underrepresented emotion categories. The proposed enhancements aim to optimize the model's robustness and generalization, ensuring effective emotion recognition across diverse and imbalanced datasets.

This section presents the proposed Emotion-Aware RoBERTa architecture, which consists of five main components: a preprocessing block, the TF-IDF-based gating mechanism, the RoBERTa backbone, the ESA layer, and the emotion prediction head. Together, these components form an end-to-end pipeline for robust emotion recognition. Before a detailed analysis of these components, a justification is provided for selecting RoBERTa as the baseline model to contextualize its suitability for the proposed approach.

Baseline model selection

The choice of RoBERTa as the baseline model in this study is motivated by its strong contextual representation capabilities, which make it well suited for emotion recognition tasks where capturing subtle emotional signals and dependencies at the phrase level is critical. RoBERTa refines the original BERT architecture by eliminating the Next Sentence Prediction (NSP) objective, introducing dynamic masking, and undergoing more extensive training with larger datasets[4]. These optimizations improve its ability to encode nuanced linguistic structures, improving performance in text classification and sentiment analysis tasks.

Compared to other transformer-based models, RoBERTa offers two key advantages:

- Enhanced contextual encoding: Unlike ALBERT and DistilBERT, which prioritize efficiency by reducing model parameters, RoBERTa retains full model capacity while optimizing its training strategy[30].
- Robust handling of long-range dependencies: Emotion recognition often involves analyzing relationships across multiple words, phrases, or clauses. RoBERTa's training optimizations allow it to better retain and process these dependencies[31].

While models such as DistilBERT and ALBERT offer computational efficiency, they come with trade-offs in representation capacity, potentially limiting their ability to capture nuanced emotional expressions. Given the need for fine-grained textual understanding in emotion recognition, RoBERTa was chosen as the backbone model to provide a robust foundation for the proposed enhancements. The methodological improvements outlined in the following subsections further refine the model's ability to prioritize emotionally salient information, ensuring more accurate and context-aware emotion recognition.

Preprocessing pipeline

Preprocessing plays an important role in ensuring the quality and integrity of textual input before feeding it into the RoBERTa model. The pipeline begins with text normalization, where noise elements such as URLs, special characters, emojis, and excessive whitespace are removed to standardize the input. Additionally, slang-aware tokenization is applied to effectively handle informal expressions, abbreviations, and domain-specific terminology, particularly in social media text. User mentions and non-alphanumeric symbols are filtered out to prevent irrelevant tokens from influencing the model's learning process.

To address class imbalance, random oversampling is employed to increase the representation of minority emotion classes such as *surprise* and *fear*. Unlike synthetic data augmentation techniques, this approach maintains the original linguistic structure of underrepresented categories, reducing bias while improving model generalization. Additionally, a TF-IDF based feature selection process is applied to pre-filter low-information words before tokenization. This step serves as a key preprocessing phase of the broader TF-IDF based gating mechanism, ensuring that only emotionally salient words remain in the dataset.

TF-IDF based gating mechanism

The TF-IDF based gating mechanism is introduced to refine attention distribution, enhancing the model's ability to focus on the most informative words when computing attention scores. This is achieved by computing TF-IDF scores after oversampling, filtering out words with low significance using a threshold of 3.5. This threshold was empirically determined through validation set tuning, where several values between 2.0 and 4.0 were evaluated. A threshold of 3.5 was found to offer the best trade-off between noise reduction and the retention of emotionally meaningful tokens. To further mitigate data sparsity, a minimum of four tokens per sample is preserved. This process effectively reduces irrelevant noise and improves emotional feature representation, thereby strengthening the model's capacity to distinguish nuanced emotional expressions.

Unlike conventional feature selection techniques that entirely remove low TF-IDF words during preprocessing, the TF-IDF based gating mechanism modulates token embeddings before passing them into RoBERTa. Each token's contribution is dynamically adjusted based on its informativeness, allowing words with high emotional relevance to receive greater emphasis while filtering out less significant ones. By integrating TF-IDF based gating with the ESA layer, the model enhances its ability to capture nuanced emotional expressions, thereby improving classification accuracy across diverse datasets.

Beyond improving classification performance, this mechanism also plays a crucial role in reducing computational overhead. RoBERTa is known for its high time complexity due to extensive self-attention computations[32]. The TF-IDF based gating mechanism alleviates this issue by filtering out less informative tokens, reducing the number of processed tokens, and consequently optimizing the model's inference time.

RoBERTa backbone

The pretrained RoBERTa[4] serves as the backbone encoder in the proposed architecture, responsible for generating deep contextualized embeddings from input text. Specifically, RoBERTa processes batches of tokenized input sequences and outputs a tensor of contextual embeddings denoted as $E \in \mathbb{R}^{B \times L \times H}$, where:

- B: Batch size, representing the number of input sequences processed simultaneously.
- L: Sequence length, the number of tokens in each input sequence.
- H: Hidden dimension, indicating the size of each token embedding.

These embeddings capture both the syntactic and semantic context of each token in the input sequence, serving as the foundation for the downstream layers, including the ESA layer, to ultimately perform emotion-aware classification.

Emotion-specific attention (ESA) layer

The ESA layer introduces a novel mechanism that extends the standard self-attention framework by incorporating emotion-aware weighting. Unlike conventional attention mechanisms that assign weights based solely on contextual relevance, the ESA layer dynamically emphasizes tokens that are emotionally salient. This is achieved by learning an emotion-specific importance vector that modulates the attention distribution, thereby amplifying the contribution of emotionally charged words while suppressing neutral or less relevant tokens. By explicitly reweighting the attention scores toward emotion-bearing tokens, ESA enhances the model's sensitivity to nuanced emotional expressions such as sarcasm, frustration, or ambivalence, emotions that are typically difficult to capture using traditional attention mechanisms[27]. This focused refinement improves the model's capacity

for fine-grained emotion classification, enabling it to differentiate between closely related emotions and better handle context-dependent sentiment cues.

The process begins by incorporating positional encodings $P \in \mathbb{R}^{L \times H}$ to retain information about the order of tokens in the sequence:

$$E_{\text{input}} = E + P \tag{1}$$

- P represents the positional encodings that inject sequence-level positional information into the token embeddings, enabling the model to distinguish the relative positions of tokens within each input sequence.

Next, the input embeddings E_{input} are linearly projected into query (Q), key (K), and value (V) matrices using trainable weight matrices $W_q, W_k, W_v \in \mathbb{R}^{H \times H}$:

$$Q = E_{\text{input}} W_q, \quad K = E_{\text{input}} W_k, \quad V = E_{\text{input}} W_v \tag{2}$$

where

- Q: Encodes the current token's request for information from other tokens.
- K: Represents the contextual features of all tokens used to match against the query.
- V: Holds the actual information to be weighted and aggregated based on attention scores.

This transformation allows each token to interact with all others in the sequence, enabling the model to compute attention scores that determine which tokens are most relevant for capturing emotional context.

Subsequently, attention scores $A \in \mathbb{R}^{L \times L}$ are calculated using the scaled dot-product attention mechanism:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{H}}\right) \tag{3}$$

where

- $QK^\top$:Computes the pairwise dot product between queries and keys, capturing token-to-token relevance.
- $\sqrt{H}$: A scaling factor that prevents large dot-product values, ensuring more stable gradients during training.
- $\text{softmax}(\cdot)$: Applies normalization along each row to convert raw attention scores into a probability distribution over the input sequence.

This step allows each token to selectively attend to other tokens based on contextual relevance, forming the foundation of the self-attention mechanism.

To prevent padding tokens from influencing the attention mechanism, an attention mask is applied. This mask assigns negligible weights to padded positions, ensuring that attention computations are only influenced by meaningful content. Using the computed attention weights, a weighted sum of the value vectors is performed, yielding the contextualized output embeddings $Z \in \mathbb{R}^{B \times L \times H}$, where tokens deemed more relevant, especially in terms of emotional content, receive greater emphasis.

At this stage, the ESA mechanism diverges from standard attention by introducing an additional emotion-aware enhancement. Specifically, a learnable scaling vector $S \in \mathbb{R}^{H}$is applied element-wise to Z, dynamically amplifying features that are emotionally salient. Unlike conventional attention mechanisms that treat contextual relevance uniformly, ESA adjusts these representations based on the emotional importance of the tokens. This dynamic reweighting ensures that subtle emotional cues, such as frustration, sarcasm, or subdued joy, are not overshadowed by dominant contextual features.

To maintain temporal coherence, positional encodings are re-applied to the scaled output embeddings, reinforcing the sequential structure of the input and ensuring that token order remains an integral part of the model's final representation. The enriched embeddings, now carrying both contextual and emotion-specific saliency, are passed to the classification head for final emotion prediction. This enables the model to perform fine-grained emotion recognition, even in the presence of nuanced or overlapping emotional cues.

Accordingly, the overall operation of the ESA mechanism can be summarized in two main steps:

1. Contextual Attention Calculation (identical to standard attention)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{H}}\right) V \tag{4}$$

2. Emotion-Specific Refinement and Sequence Restoration:

$$Z_{\text{final}} = (\text{Attention}(Q, K, V) \odot S) + P \tag{5}$$

where

- $\odot$ denotes element-wise multiplication with the learned scaling vector $S \in \mathbb{R}^H$, emphasizing emotionally significant features,
- $P \in \mathbb{R}^{L \times H}$ is the positional encoding matrix reintroduced to preserve order information.

This formulation distinguishes the attention of ESA from the standard attention by integrating a learnable emotion-awareness component that selectively amplifies emotionally relevant patterns, thereby improving the model's interpretability and performance in emotion recognition tasks.

Emotion prediction head

The emotion prediction head is responsible for translating the refined contextual embeddings into emotion predictions. It consists of a fully connected (dense) layer followed by a softmax activation function, which produces a probability distribution over the predefined emotion classes. This design enables the transformation of high-dimensional hidden states into a lower-dimensional space tailored for emotion classification tasks[33], supporting practical applications such as sentiment analysis, mental health monitoring, and content moderation.

Figure 1 illustrates the end-to-end architecture of the proposed Emotion-Aware RoBERTa framework. The process begins with a preprocessing pipeline that performs text cleaning, tokenization, and normalization to prepare the input for the RoBERTa encoder. RoBERTa then generates contextualized embeddings using its multi-head self-attention mechanism, capturing deep semantic relationships between tokens. These embeddings are subsequently processed by the ESA layer, which re-weights the attention distribution to prioritize emotionally salient tokens. The figure visually demonstrates how ESA emphasizes key emotional expressions within the input sequence while down-weighting irrelevant content. The refined outputs are then pooled and passed through one or more fully connected layers, culminating in the softmax layer that outputs the final emotion class probabilities.

The following section presents a comprehensive evaluation of the proposed model, including an ablation study and performance analysis. The results demonstrate how the integration of the ESA layer, in conjunction

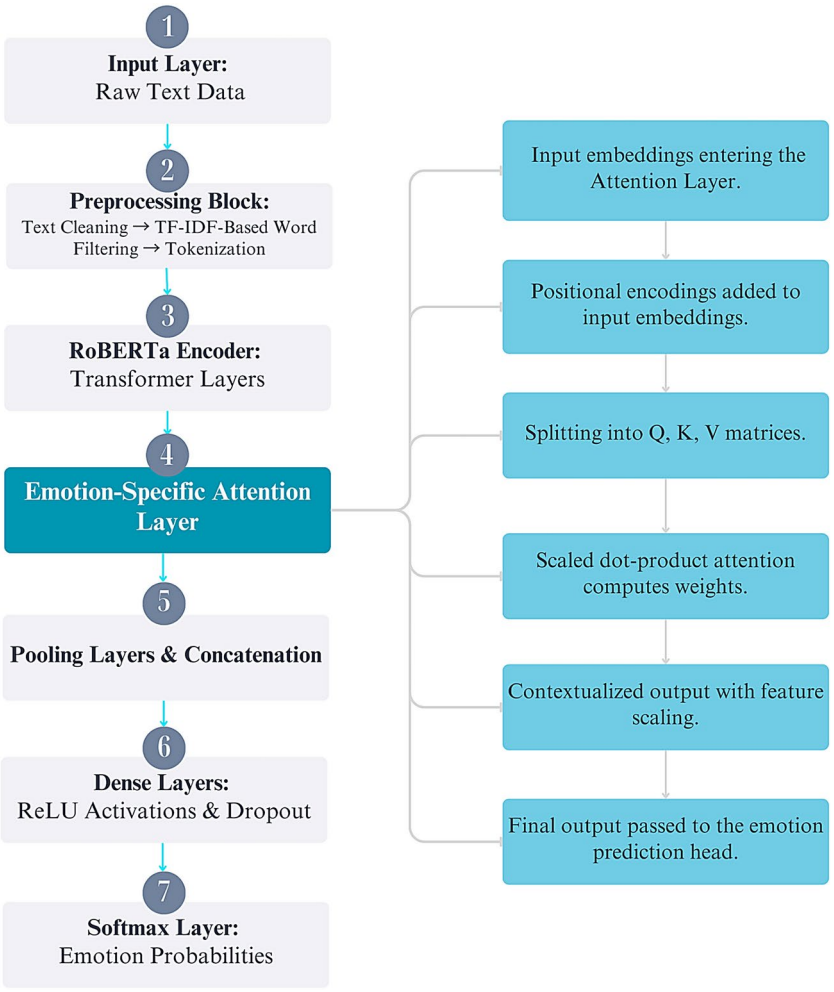

Fig. 1. Architecture of the proposed Emotion-Aware RoBERTa framework. The system integrates preprocessing, TF-IDF-based gating, the RoBERTa encoder, the ESA layer, and the classification head to emphasize emotionally salient tokens for improved emotion recognition.

with the preprocessing and TF-IDF based gating mechanisms, contributes to improved emotion recognition by enhancing focus on emotionally relevant content and reducing noise sensitivity.

Datasets and experimental setup

Primary dataset description

The primary dataset utilized in this study was obtained from Kaggle's "Emotions Dataset for NLP" repository[34]. It comprises 21000 text samples categorized into six distinct emotion classes: *joy*, *sadness*, *anger*, *fear*, *love*, and *surprise*. These categories encompass a wide range of emotional expressions, making the dataset a valuable resource for evaluating emotion recognition models. Notably, the dataset exhibits a significant class imbalance, as illustrated in Table 1, with "joy" and "sadness" appearing more frequently, whereas "surprise" and "fear" are underrepresented. This imbalance poses challenges in developing a model that generalizes well across all emotion classes.

Additionally, the dataset captures a range of linguistic nuances commonly observed in informal communication, such as:

- Sarcasm Indicators: Sentences may contain explicit indicators of sarcasm, such as exaggerated phrases (e.g., "*Oh, that's just perfect*!") or punctuation patterns that amplify tone. These subtle cues require the model to consider contextual meaning rather than just token-level semantics.
- Idiomatic Expressions: The dataset includes emotionally charged idioms and phrases, such as "*walking on air*" (*joy*) or "*in the depths of despair*" (*sadness*). Correctly interpreting these idioms necessitates contextual understanding.
- Slang and Abbreviations: Informal text often includes slang terms ("*kinda*," "*gonna*") and abbreviations, posing additional challenges during preprocessing and modeling.

Data preprocessing

The preprocessing pipeline was meticulously designed to ensure the dataset was optimally prepared for training, incorporating the following steps:

- Text cleaning: Non-informative characters, excessive whitespace, and special symbols were removed to normalize the text. Punctuation was preserved where it contributed to the emotional context.
- Tokenization: The RoBERTa tokenizer was employed to segment sentences into subword units, ensuring effective handling of out-of-vocabulary words and maintaining contextual integrity[35].
- Class balancing: Given the inherent class imbalance, a random oversampling technique was applied to augment underrepresented categories, such as fear and surprise, by duplicating existing samples. Unlike synthetic resampling methods like SMOTE, this approach retains the dataset's original distribution while enhancing the model's ability to generalize across all emotion classes[36].

Experimental setup

The experimental setup was carefully designed to ensure efficient training, evaluation, and reproducibility of the proposed approach.

Platform and tools

- Computational environment: Experiments were conducted on Google Colab Pro+, utilizing an NVIDIA A100-SXM4-40GB GPU to accelerate training and evaluation. The environment was set up with Python 3.11.11 and TensorFlow 2.18.0, running on a cloud instance equipped with approximately 89.6 GB of RAM. For code development and initial testing, a local Windows 11 system was used, featuring an Intel Core Ultra 9 185H processor, 16 GB of RAM, and an Intel Arc Graphics GPU.
- ***Key libraries:***
  - Hugging Face Transformers: Used for loading and fine-tuning RoBERTa[37].
  - PyTorch: Employed for model implementation, optimization, and training loop management.
  - Imbalanced-Learn: Applied for handling class imbalance through oversampling techniques.
  - Matplotlib & Seaborn: Utilized for visualizing training dynamics and generating confusion matrices.
  - NLTK: Used for preprocessing tasks, including tokenization and text normalization.

Emotion class	Number of samples	Percentage (%)
Joy	6761	33.81
Sadness	5794	28.97
Anger	2709	13.55
Fear	2373	11.87
Love	1641	8.21
Surprise	719	3.60
Total	20,997	100.00

Table 1. Dataset distribution across emotion classes.

Dataset splitting strategy
To ensure robust model training and evaluation, the dataset was partitioned as follows:

- Training set (80%): Used for model learning and parameter optimization.
- Validation set (10%): Employed to monitor model performance during training and fine-tune hyperparameters.
- Testing set (10%): Held out for final performance evaluation on unseen data.

Training procedure
The model was trained for 10 epochs with a batch size of 16, ensuring a balance between training speed and GPU memory efficiency. Cross-entropy loss was employed for multi-class classification, while the AdamW optimizer was utilized for weight updates. To enhance training stability, a learning rate scheduler was integrated to dynamically adjust the learning rate throughout the training process.

Hyperparameter tuning
Hyperparameter tuning was conducted to optimize the model's performance while maintaining computational efficiency. The values in Table 2 were selected through iterative experimentation to balance performance and training stability. A batch size of 16 was chosen to fit the GPU memory while maintaining training efficiency. The learning rate (1×10^{-5}) was optimal for stable fine-tuning of RoBERTa, preventing overfitting. Ten epochs ensured convergence without degrading performance. A dropout rate of 0.3 helped to reduce overfitting and to regularize the classification head, especially under class imbalance. Weight decay (0.01) improved generalization, and a learning rate scheduler was used to stabilize early training. These values were refined through grid search and validation experiments, yielding the best accuracy and F1-score across datasets.

To ensure optimal performance, hyperparameters such as learning rate, batch size, and number of epochs were selected based on preliminary tuning using the validation dataset. The learning rate of 1×10^{-5}, a batch size of 16, and 10 training epochs consistently yielded strong results across multiple test runs. During sensitivity analysis, minor variations in the learning rate (between 5×10^{-6} and 2×10^{-5}) and number of epochs (ranging from 8 to 12) showed negligible impact (less than 0.5%) on both accuracy and F1-score, confirming the stability and robustness of the chosen settings. These selections balance training efficiency and generalization performance, making them suitable for real-world applications.

Evaluation metrics
To comprehensively assess the model's performance, several evaluation metrics were employed:

- Accuracy: Measures the overall correctness of the model by calculating the proportion of correctly predicted instances over the total number of samples.
- Precision, Recall, and F1-Score: These metrics were computed for each emotion class to provide a class-wise evaluation. Precision evaluates the proportion of true positive predictions among all predicted positives, recall assesses the proportion of true positives captured among all actual positives, and the F1-Score represents the harmonic mean of precision and recall. These metrics are particularly important in the presence of class imbalance.
- Confusion Matrix: Provides a detailed breakdown of true versus predicted labels across emotion categories, enabling the identification of common misclassifications and patterns of confusion, particularly between semantically similar emotions.
- Matthews Correlation Coefficient (MCC): Used as a robust indicator of classification performance, especially for imbalanced datasets. MCC takes into account true and false positives and negatives and returns a value between -1 and 1, where 1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation.

Rationale for choosing MCC: The MCC was selected over Cohen Kappa for two reasons[38]:

- First, MCC incorporates all four components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a balanced and reliable evaluation even when class distributions are skewed. This makes it particularly useful for emotion recognition tasks where certain emotions are underrepresented.
- Second, since a weighted F1-score is already used to capture class-level performance, MCC serves as a complementary global measure that evaluates the model's overall ability to correctly classify across all categories.

Hyperparameter	Value
Learning rate	1×10^{-5}
Batch size	16
Dropout rate	0.3
Number of epochs	10
Optimizer	AdamW

Table 2. Hyperparameter configuration.

Emotion	Baseline RoBERTa			Emotion-Aware RoBERTa			Improvement in F1-score (%)
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Anger	0.90	0.84	0.86	0.97	0.99	0.98	+12.00
Fear	0.80	0.82	0.81	0.97	0.98	0.97	+16.00
Joy	0.93	0.84	0.88	0.96	0.90	0.93	+5.00
Love	0.72	0.88	0.79	0.96	0.99	0.98	+19.00
Sadness	0.84	0.92	0.88	0.96	0.94	0.95	+7.00
Surprise	0.94	0.72	0.82	0.98	1.00	0.99	+17.00
Accuracy	85.90			96.77			+10.87
Weighted F1-score	0.85	0.70	0.84	0.96	0.97	0.97	+13.00

Table 3. Comparison of classification performance: baseline RoBERTa vs. Emotion-Aware RoBERTa (ESA with TF-IDF).

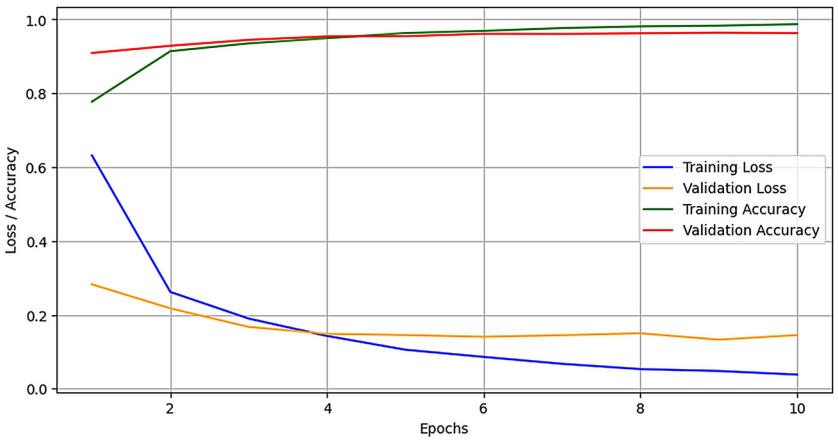

Fig. 2. Training and validation metrics of Emotion-Aware RoBERTa. The steady reduction in loss and stability in accuracy indicate successful convergence and effective generalization.

While the weighted F1-score emphasizes performance per class, MCC offers a holistic view of classification balance.

Experimental results and discussion

This section presents a comprehensive evaluation of the Emotion-Aware RoBERTa model. The evaluation includes a comparison against the baseline RoBERTa[4] to assess the effectiveness of the proposed enhancements. An ablation study is also conducted to isolate the impact of the ESA layer by evaluating the model without the TF-IDF gating mechanism. In addition, the model's performance is compared with other transformer-based models, including ALBERT[14] and DistilBERT[15]. Finally, the generalization capability of the Emotion-Aware RoBERTa is examined using external datasets, offering insights into its robustness across diverse linguistic structures and domain variations.

Performance evaluation of the emotion-aware RoBERTa

The Emotion-Aware RoBERTa demonstrated outstanding performance on the primary dataset, demonstrating its ability to effectively classify emotions across the aforementioned categories. As reported in Table 3, the model achieved an overall accuracy of 96.77% and a weighted F1-Score of 0.97. It is demonstrated that an F1-Score exceeding 0.9 (or 90%) indicates near-perfect classification with minimal misclassification, a performance level rarely achieved in complex NLP tasks[39].

Remarkably, minority classes such as *surprise* and *fear*, which are often difficult for traditional models, achieved near-perfect F1-Scores of 0.99 and 0.97, respectively. The model also performed exceptionally well in majority classes like *joy* and *sadness*, attaining F1-Scores of 0.93 and 0.95, respectively. These significant improvements can be attributed to the ESA layer's capability to focus on emotionally important tokens while preserving strong contextual representations.

Figure 2 illustrates the model's optimization process, highlighting a steady decline in both training and validation loss across epochs. This consistent reduction, coupled with sustained high accuracy, indicates an effective learning process. The model converges by the fifth epoch with minimal overfitting, maintaining stable performance in subsequent epochs. These results demonstrate the robustness and reliability of the proposed training strategy, ensuring generalizability across different datasets, as will be shown in the subsequent analysis.

As an additional metric, the confusion matrix in Fig. 3 provides further validation of the model's performance. The majority of classifications are correctly placed along the diagonal, with minimal misclassifications, particularly between closely related emotions like *joy* and *love*.

Although the Emotion-Aware RoBERTa exhibited outstanding overall performance, an analysis of misclassified instances reveals areas for potential improvement. Most errors occurred in cases where linguistic patterns overlapped, particularly between semantically similar emotions such as *joy* and *love*. For example, the sentence *"Love everything learning feel really passionate design"* was misclassified as *love* instead of *joy*, highlighting the challenge of distinguishing closely related positive emotions. Another instance is *"Being around friends makes me feel safe and warm"*, which was predicted as *joy* while the ground truth was *love*, likely due to the subtle affective overlap in the expression of interpersonal warmth. Similarly, the sentence *"I'm grateful for this opportunity, but also nervous about what's ahead"* was predicted as *joy* instead of *fear*, where co-occurrence of mixed sentiments confuses the model.

These examples illustrate that despite ESA's capability to emphasize emotionally salient tokens, it may still struggle to disambiguate fine-grained emotional contexts when lexical cues are shared across classes. Additionally, longer sentences containing multiple emotional nuances posed difficulties, as the model often prioritized the dominant emotion while underrepresenting subtler emotional undertones. Future refinements to ESA could incorporate polarity-aware attention mechanisms or leverage external semantic knowledge to more effectively separate closely related emotion categories. Despite these challenges, the overall misclassification rate remained lower than that of the baseline RoBERTa, reinforcing the effectiveness of the proposed enhancements in improving fine-grained emotion recognition.

Performance comparison between baseline and Emotion-Aware RoBERTa

The Emotion-Aware RoBERTa demonstrated substantial improvements over the baseline RoBERTa, as summarized in Table 3. The proposed enhancements led to an increase in overall accuracy from 85.90% to 96.77%, representing an improvement of approximately 11%. Additionally, the weighted F1-Score improved from 0.84 to 0.97, underscoring the model's enhanced capability to classify emotions across both majority and minority classes effectively. The impact of these enhancements was particularly evident in minority classes such as *surprise* and *fear*. For example, the baseline model achieved a recall of 0.72 for *surprise*, whereas the Emotion-Aware RoBERTa significantly improved this to 1.00. This improvement highlights the effectiveness of the ESA layer in prioritizing key emotional cues, thereby enhancing the model's ability to classify underrepresented emotions with greater accuracy.

Additionally, the Emotion-Aware RoBERTa effectively reduced misclassification errors in closely related emotion classes such as *joy* and *love*, which the baseline model frequently misclassified due to overlapping linguistic patterns. The improved differentiation between these nuanced emotions demonstrates the robustness of the proposed architecture in capturing fine-grained emotional expressions with higher precision. To further understand the internal behavior and contributions of the proposed enhancements, this section delves deeper into the specific roles played by the ESA layer and the TF-IDF based gating mechanism. Beyond overall performance gains, it is essential to examine how these components influence attention distribution, reduce noise, and improve inference efficiency. Both visualization and ablation studies are presented to highlight their individual and combined effectiveness.

Effectiveness of the ESA layer and impact of TF-IDF gating

The integration of the ESA layer with the TF-IDF based gating mechanism plays a pivotal role in enhancing the proposed model's ability to focus on emotionally salient information while filtering out noise. To assess

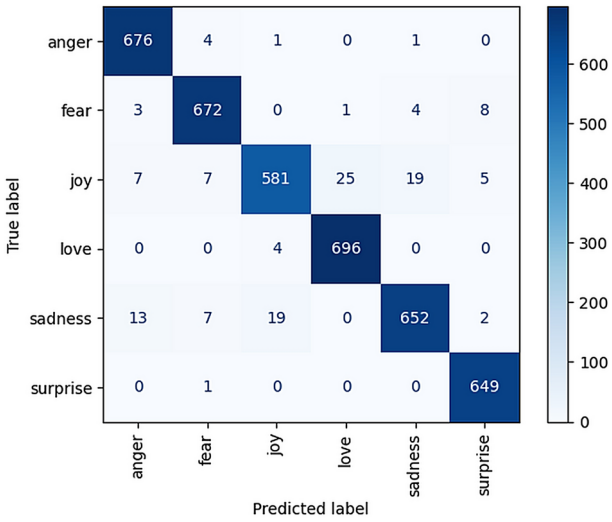

Fig. 3. Confusion matrix of Emotion-Aware RoBERTa on the primary dataset, showing accurate classification across six emotion categories with minimal misclassification, especially between nuanced emotions.

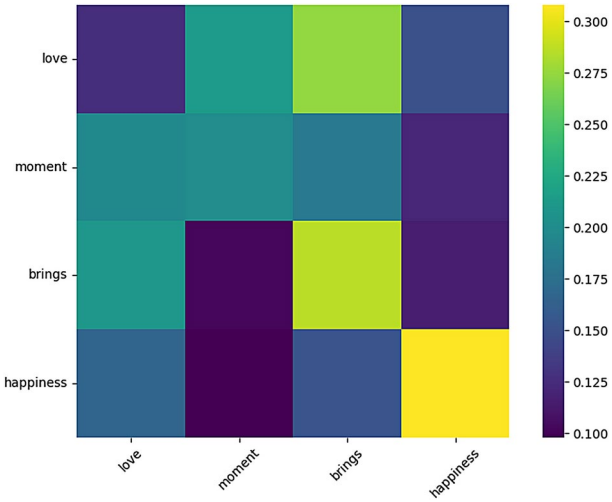

Fig. 4. Attention heatmap of Emotion-Aware RoBERTa for the sentence "*Love this moment, it brings happiness*," highlighting the model's focus on emotionally salient tokens such as "*love*" and "*happiness*.".

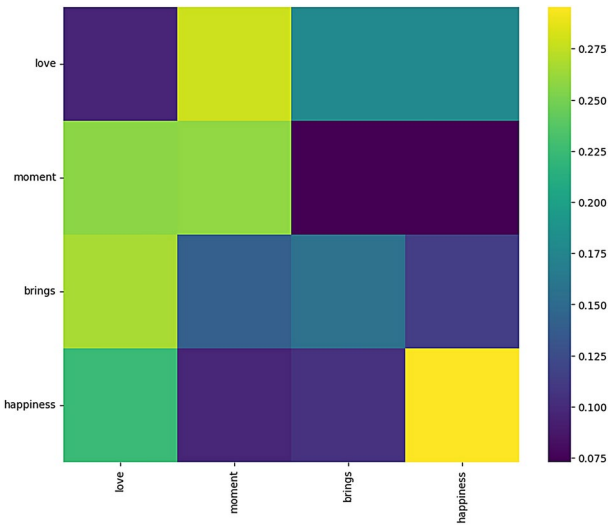

Fig. 5. Attention heatmap of baseline RoBERTa for the sentence "*Love this moment, it brings happiness*," showing a more distributed focus across the sentence compared to the Emotion-Aware RoBERTa, with less emphasis on the emotionally salient tokens "*love*" and "*happiness*.".

the contribution of this combination, both qualitative and quantitative analyses were conducted. A qualitative comparison of attention distributions was performed using the sentence *"Love this moment, it brings happiness."* As illustrated in Fig. 4, the Emotion-Aware RoBERTa assigns high attention weights to emotionally charged tokens such as *love* and *happiness*, while down-weighting less informative words and moderately attending to relevant context like *brings*. In contrast, the baseline RoBERTa, as shown in Fig. 5, spreads attention more uniformly, including tokens that are less emotionally informative.

This behavior is further supported by the attention scores reported in Tables 4 and 5. The Emotion-Aware RoBERTa consistently attributes greater weight to key emotional expressions. The word-level saliency maps in Figs. 6 and 7 offer an additional visualization of the differential focus, clearly showing how ESA and TF-IDF gating jointly enhance interpretability and emotional relevance.

To isolate the specific impact of the TF-IDF gating mechanism, an ablation study was conducted. A variant of the proposed model was trained using only the ESA layer, excluding the TF-IDF component. As reported in Table 6, this configuration achieved a slightly higher accuracy (98.35%) than the full model (96.77%) presented in Table 6. However, this modest gain came at the expense of increased inference time and reduced robustness to noisy or imbalanced input. This can be attributed to longer input sequences containing more low-information tokens, which the TF-IDF gating mechanism would otherwise suppress.

In contrast, incorporating TF-IDF gating reduced uninformative content, leading to shorter sequences, faster inference, and greater stability, particularly in real-world applications where text often includes slang,

Index	Token	Attention score
0	Love	0.1906
1	Moment	0.1750
2	Brings	0.1783
3	Happiness	0.1813

Table 4. Emotion-Aware RoBERTa word-level importance scores.

Index	Token	Attention score
0	Love	0.1844
1	Moment	0.1784
2	Brings	0.1687
3	Happiness	0.1650

Table 5. Baseline RoBERTa word-level importance scores.

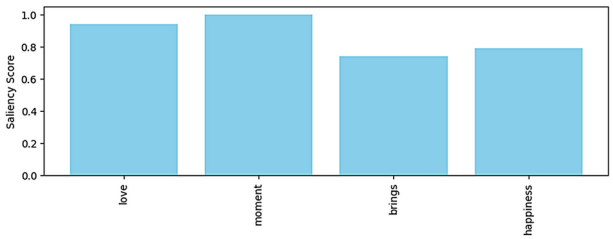

Fig. 6. Token-level saliency map from Emotion-Aware RoBERTa for the sentence "Love this moment, it brings happiness." Emotionally salient words receive higher importance scores.

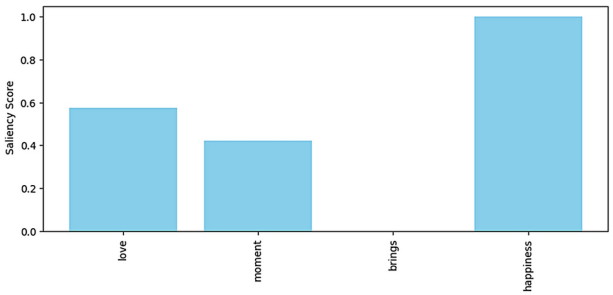

Fig. 7. Token-level saliency map from baseline RoBERTa for the same sentence. The attention is less concentrated on emotionally salient words, reflecting the absence of emotion-specific refinement.

Emotion	Baseline RoBERTa			Emotion-Aware RoBERTa			Improvement in F1-score (%)
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Anger	0.93	0.93	0.93	0.98	0.99	0.99	+6.00
Fear	0.88	0.90	0.89	0.98	0.99	0.98	+9.00
Joy	0.98	0.92	0.95	0.99	0.94	0.97	+2.00
Love	0.78	0.96	0.86	0.96	0.99	0.98	+12.00
Sadness	0.96	0.98	0.97	1.00	0.98	0.99	+2.00
Surprise	0.89	0.72	0.80	0.98	1.00	0.99	+19.00
Accuracy	93.10			98.35			+5.25
Weighted avg	0.93	0.92	0.93	0.98	0.98	0.98	+5.0

Table 6. Comparison of classification performance: baseline RoBERTa vs. enhanced RoBERTa (ESA without TF-IDF).

Metric	RoBERTa	DistilBERT	ALBERT	Emotion-Aware RoBERTa
Accuracy	85.90	87.10	85.85	**96.77**
F1-score (weighted)	0.84	0.87	0.86	**0.97**
Matthews correlation coefficient (MCC)	0.82	0.82	0.81	**0.96**
Training time (minutes)	04	**02**	03	08
Memory usage	Moderate	Moderate	**Low**	High
Misclassification rate	Low	Moderate	High	**Lowest**

Table 7. Comparison of baseline models vs. Emotion-Aware RoBERTa on the primary dataset.

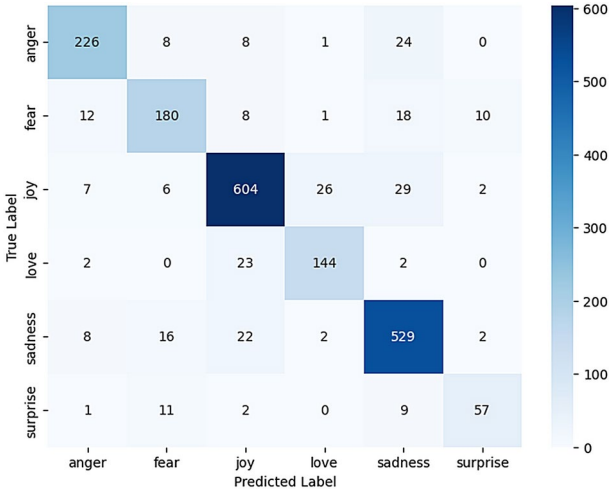

Fig. 8. Confusion matrix for DistilBERT, revealing higher misclassification rates for closely related emotions, such as "*joy*" and "*fear*," compared to Emotion-Aware RoBERTa.

typos, and informal expressions. This trade-off highlights the practical value of combining ESA and TF-IDF, offering a balance between high classification accuracy and computational efficiency. Overall, the ESA layer provides a targeted refinement of attention based on emotional salience, while the TF-IDF gating mechanism improves efficiency and noise robustness. Together, they enable Emotion-Aware RoBERTa to achieve superior performance, interpretability, and scalability in emotion classification tasks.

Comparison with benchmarks: DistilBERT and ALBERT

The aforementioned findings confirm that the integration of ESA and TF-IDF-based gating not only enhances classification accuracy and attention focus but also improves the model's robustness in handling imbalanced and noisy text, while maintaining computational efficiency. To further contextualize the strength of the proposed architecture, this section presents a comparative evaluation against other transformer-based models, namely DistilBERT and ALBERT, to demonstrate the competitiveness of the Emotion-Aware RoBERTa across broader benchmark scenarios.

Table 7 presents a comparative analysis of the baseline and enhanced versions of RoBERTa alongside these widely used transformer-based models. All models were trained and evaluated under identical experimental settings using the same primary dataset to ensure a fair and consistent comparison. The results show that the Emotion-Aware RoBERTa outperformed all other models, achieving an accuracy of 96.77% and a weighted F1-Score of 0.97. In contrast, DistilBERT achieved an accuracy of 87.10% with a matching weighted F1-Score of 0.97, while ALBERT reached 85.85% accuracy and a weighted F1-score of 0.86. Although DistilBERT outperformed ALBERT in both metrics, the Emotion-Aware RoBERTa consistently demonstrated superior classification performance, validating the effectiveness of the proposed enhancements.

In addition to accuracy and F1-Score, Table 7 also reports the Matthews Correlation Coefficient (MCC), a robust metric particularly suited for evaluating imbalanced classification tasks. The results show that the Emotion-Aware RoBERTa achieved the highest MCC score among all compared models, further affirming its effectiveness in minimizing classification errors and maintaining strong generalization across diverse emotion categories.

Furthermore, misclassification patterns were examined to gain deeper insights into model behavior. The Emotion-Aware RoBERTa exhibited the lowest misclassification rate, particularly in semantically similar emotion pairs such as joy and fear. Figures 8 and 9, which depict the confusion matrices for DistilBERT and ALBERT, respectively, highlight their challenges in distinguishing between overlapping emotional categories.

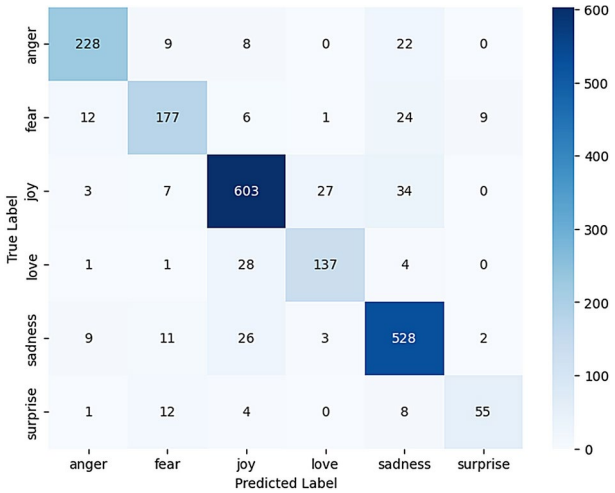

Fig. 9. Confusion matrix for ALBERT, highlighting increased misclassification errors, particularly for nuanced emotions, attributed to its compressed architecture.

While both DistilBERT and ALBERT demonstrated lower classification performance, they offer benefits in computational and memory efficiency.

This analysis underscores the practical trade-offs between accuracy, F1-score, MCC, misclassification rate, and computational cost, offering valuable insights for selecting models based on application-specific requirements, such as deployment constraints or the importance of fine-grained emotional resolution.

Statistical significance analysis

To assess whether the Emotion-Aware RoBERTa significantly outperforms the baseline RoBERTa with TF-IDF gating, McNemar's test, a non-parametric method for comparing paired nominal data, was employed. This test is specifically designed to evaluate whether two models differ significantly in their predictions on the same dataset by focusing on instances where the models disagree in their classifications[40].

Unlike conventional accuracy metrics, McNemar's test highlights whether performance differences are statistically meaningful, especially in terms of prediction behavior. The test revealed a highly significant difference ($p < 0.001$), confirming that Emotion-Aware RoBERTa makes substantially different, and more accurate, predictions than the baseline. Based on the observed disagreement counts, the McNemar's test statistic exceeded the critical threshold for $p < 0.001$, providing strong statistical evidence of improvement.

Notably, Emotion-Aware RoBERTa correctly predicted samples that the baseline misclassified in roughly 7.5% of test cases, highlighting its enhanced ability to capture emotional nuance. For example:

Text: feel like i'm assaulted

Ground Truth: fear

Baseline Prediction: sadness

Emotion-Aware RoBERTa Prediction: fear

This demonstrates ESA's strength in identifying subtle emotions more accurately. Overall, McNemar's test supports that Emotion-Aware RoBERTa not only performs better, but does so in a meaningfully different way. This example illustrates how the proposed model, through the ESA layer and TF-IDF gating, effectively distinguishes fine-grained emotional expressions. Overall, McNemar's test validates that the performance gain is not only empirical but also statistically significant, reinforcing the reliability and generalizability of the proposed enhancements.

Computational efficiency and deployment feasibility

To ensure the computational efficiency and feasibility of the Enhanced RoBERTa for real-world deployment, inference-time optimization was performed using Automatic Mixed Precision (AMP) and FP16 computation. These optimizations were selected to accelerate inference speed and minimize GPU memory usage without compromising classification accuracy, making the model suitable for large-scale applications. In the experiments of this paper, the FP16 + AMP optimization was applied during the evaluation phase, following model training. Specifically, the model was converted to half-precision (FP16), and the inference was conducted using PyTorch's `torch.amp.autocast()` to enable mixed-precision execution. This approach was preferred over full quantization, which may introduce performance degradation due to reduced numerical precision[41]. Table 8 summarizes the impact of these techniques on inference efficiency.

The optimization was applied at the inference stage, meaning that the model underwent FP16 conversion after training and validation were completed. The following steps were implemented:

Optimization step	Total inference time (S)	Avg inference time per sample (S)	Max GPU memory used (MB)
Before FP16 + AMP	1.4451	0.000356	2818.54
After FP16 + AMP	1.1620	0.000286	1962.83

Table 8. Inference performance before and after FP16 + AMP optimization.

Dataset	No. of instances	Accuracy	Weighted F1-score
1st Dataset[42]	422,746	88.03	0.88
2nd Dataset[43]	15,574	65.67	0.70

Table 9. Generalization performance of Emotion-Aware RoBERTa on external datasets.

1. Loading the trained model: The best-performing model checkpoint was loaded.
2. FP16 conversion: The model was explicitly converted to FP16 precision using `model.to(torch.float16)`, ensuring reduced memory usage.
3. AMP-enabled inference: During inference, the computations were run within `torch.amp.autocast("cuda")`, allowing PyTorch to automatically select lower-precision operations where beneficial while maintaining full precision where necessary.
4. Memory and time benchmarking: The inference process was timed, and peak GPU memory consumption was recorded to assess efficiency improvements.

Despite the total execution time remaining approximately 8 minutes, applying FP16 + AMP resulted in a 19.6% reduction in inference time per sample and a 30.3% reduction in peak GPU memory consumption. These improvements make the Emotion-Aware RoBERTa significantly more efficient for real-world deployment on GPU-based environments, particularly for large-scale applications requiring high-throughput emotion classification.

The application of FP16 + AMP led to measurable improvements in inference efficiency, significantly reducing per-sample inference time and GPU memory consumption. These optimizations enhance the feasibility of deploying the model on real-world GPU environments, ensuring faster and more resource-efficient processing. By implementing FP16 + AMP at the inference stage, the model now achieves a balance between computational efficiency and precision, enabling its use in practical, resource-constrained environments without sacrificing classification accuracy.

Generalization across external datasets

Beyond evaluating the Emotion-Aware RoBERTa on the primary dataset, its performance was further assessed on two external datasets to examine its generalization capability[11], as summarized in Table 9. This analysis investigates how effectively the model adapts to unseen data from diverse sources and whether the enhancements introduced in the primary dataset translate to other real-world scenarios. The two external datasets present distinct challenges, including linguistic diversity, varying degrees of noise, and significant class imbalance, making them valuable benchmarks for assessing the model's robustness in handling domain shifts and generalizing beyond the training distribution.

The first external dataset, titled *Sentiment and Emotion Analysis Dataset*[42], is significantly larger, containing 422,746 records with an imbalanced distribution across six emotion labels. This dataset presents a diverse range of linguistic expressions, making it a valuable benchmark for evaluating the model's scalability. The Emotion-Aware RoBERTa achieved an accuracy of 88.03%, demonstrating its ability to handle large-scale data effectively while maintaining robust classification performance. During the experiments, the model exhibited strong results across all emotion categories, with F1-Scores ranging from 0.84 for *surprise* to 0.95 for *sadness*, and a weighted F1-Score of 0.89. Although the accuracy is lower compared to the primary dataset, the relatively strong F1-Scores indicate that the model successfully generalizes to large datasets despite domain shifts and class imbalances.

The second external dataset, titled *Sentiment and Emotions of Tweets* [43], consists of samples collected from social media platforms. Although the original dataset includes 11 emotion labels, the analysis was restricted to six, *anger, fear, joy, love, sadness, and surprise*, to ensure consistency with the primary dataset. This dataset presents additional challenges, including substantial class imbalance and increased text noise, with emotions such as *love* and *surprise* being notably underrepresented. As a result, the Emotion-Aware RoBERTa achieved an accuracy of 65.67% and a weighted F1-Score of 0.70.

The performance degradation, of the proposed model, observed in the second dataset can be attributed to several factors. First, the smaller dataset size inherently limits the model's ability to predict robust representations, particularly for rare emotions such as *surprise*, which had only 34 instances. The imbalance in emotion distribution likely contributed to reduced recall for underrepresented categories. Second, the informal and noisy nature of social media text, characterized by abbreviations, slang, and inconsistent grammar, poses additional challenges that the model may not have encountered during training on the primary dataset. Despite these difficulties, the Emotion-Aware RoBERTa effectively classified more frequent emotions such as *joy* and *anger*, achieving F1-Scores of 0.85 and 0.86, respectively.

The decline in performance across external datasets highlights the impact of domain shifts, class imbalance, and linguistic variability on model generalization. While the model performs well on structured and large-scale datasets, its accuracy declines in more challenging environments where data is limited and noisy. These findings emphasize the need for further improvements, such as domain-adaptive fine-tuning, data augmentation strategies, and enhanced preprocessing techniques, to ensure robust cross-domain emotion classification.

Overall, all results highlight the strong generalization of the Emotion-Aware RoBERTa model, as it maintains high performance in large and diverse external datasets. The model consistently demonstrates robustness and adaptability, effectively handling variations in linguistic patterns, class distributions, and dataset-specific challenges while delivering reliable classification outcomes.

Limitations and future directions

While the proposed Emotion-Aware RoBERTa demonstrates strong performance across several datasets, certain limitations remain. First, the model operates under a single-label classification assumption, assigning only one dominant emotion per text. This can be restrictive in real-world scenarios where texts often convey mixed or overlapping emotional states. Future work may explore extending the framework to support multi-label emotion recognition[44] to better capture complex emotional expressions. This could be achieved by modifying the prediction head to use a sigmoid activation function instead of softmax. This enables independent probability estimates for each emotion class and allows overlapping predictions.

Second, although the ESA layer improves attention toward emotionally salient tokens, challenges persist in accurately detecting subtle linguistic phenomena such as sarcasm, irony, and ambiguous emotional cues. These often require a deeper level of contextual understanding or external modalities (e.g., audio or visual signals)[45]. Incorporating context-aware or multimodal approaches (e.g., emoji interpretation or prosody from speech) could further enhance the model's capabilities. Third, while the model has been evaluated on diverse datasets, its performance on low-resource languages or code-mixed inputs remains untested.

Fourth, while the Emotion-Aware RoBERTa achieved competitive performance, its architecture is computationally intensive compared to lighter models like ALBERT or DistilBERT. Although inference optimization using FP16 and AMP significantly reduced memory and latency, further model compression and scalability strategies could be explored. Fifth, the current evaluation focuses on benchmark datasets. Future research may involve domain adaptation and testing in real-world or cross-lingual settings, where language variations and cultural factors influence emotional expression. Finally, although qualitative attention heatmaps were used for interpretation, a more comprehensive integration of explainable AI (XAI) techniques[46] such as SHAP or LIME could further enhance transparency and interpretability, an avenue we plan to explore in future work. By attributing importance scores to individual tokens, SHAP can provide complementary insights alongside the ESA attention heatmaps, offering a more granular understanding of how the model arrives at emotion predictions. By addressing these limitations, future extensions of this work can improve robustness, broaden applicability, and enhance deployment readiness for real-world emotion-aware NLP systems.

Conclusion

This paper presented Emotion-Aware RoBERTa, an enhanced emotion recognition framework that builds upon the standard RoBERTa model by integrating an Emotion-Specific Attention (ESA) layer and a TF-IDF based gating mechanism. These enhancements were designed to address key limitations in traditional transformer-based models, including limited sensitivity to emotionally salient tokens, difficulties with class imbalance, and challenges in handling noisy, informal text. Extensive experiments demonstrated that the proposed model significantly outperformed the baseline RoBERTa and other benchmark transformer models. Specifically, it achieved an accuracy of 96.77%, compared to 85.90% for baseline RoBERTa, 87.10% for DistilBERT, and 85.85% for ALBERT under identical experimental settings. The Emotion-Aware RoBERTa also achieved the highest weighted F1-Score and Matthews Correlation Coefficient (MCC), further validating its robust classification performance. To ensure interpretability and transparency, attention heatmaps were introduced to visualize how the ESA layer emphasizes emotionally significant tokens. Moreover, an ablation study quantified the individual contributions of the ESA and TF-IDF modules, revealing the trade-off between classification accuracy and inference efficiency. The model's real-world applicability was further supported by inference optimization using FP16 and AMP, reducing memory consumption and inference time. McNemar's statistical test was also employed to validate the significance of improvements. Finally, the model exhibited strong generalization across two external datasets with varied sizes and domain characteristics, reinforcing its adaptability to diverse linguistic patterns and real-world scenarios. These findings position Emotion-Aware RoBERTa as a scalable, interpretable, and high-performing solution for fine-grained emotion recognition.

Data availability

The primary and external datasets generated and analyzed during the current study are available in the *Kaggle* repository, in the following web links[34,42,43]: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp, https://www.kaggle.com/datasets/ankitkumar2635/sentiment-and-emotions-of-tweets, https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset.

Received: 20 February 2025; Accepted: 21 April 2025
Published online: 21 May 2025

References

1. Fortuna, P. & Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* https://doi.org/10.1145/3232676 (2018).
2. Maruf, A. A. et al. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access* **12**, 18416–18450 (2024).
3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, https://doi.org/10.18653/v1/N19-1423 (Association for Computational Linguistics, 2019).
4. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692 (2019).
5. Acheampong, F., Nunoo-Mensah, H. & Chen, W. Transformer models for text-based emotion detection: A review of bert-based approaches. *Artif. Intell. Rev.* **54**(54), 5789–5829 (2021).
6. Oprea, S. & Magdy, W. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2854–2859 (Association for Computational Linguistics, 2019).
7. Shu, X. Bert and roberta for sarcasm detection: Optimizing performance through advanced fine-tuning. *Appl. Comput. Eng.* **97**, 1–11 (2024).
8. Zhou, J. An evaluation of state-of-the-art large language models for sarcasm detection. *arXiv preprint* arXiv:2312.03706 (2023).
9. Poria, S., Cambria, E., Hazarika, D. & Vij, P. Affective computing for social media text analysis: A survey. *Inf. Fusion* **54**, 126–144. https://doi.org/10.1016/j.inffus.2019.07.001 (2020).
10. Sosea, R. & Caragea, C. Multimodal sarcasm detection with contextual word-based sparsity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3137–3150 https://doi.org/10.18653/v1/2021.emnlp-main.252 (Association for Computational Linguistics, 2021).
11. Budnikov, M., Bykova, A. & Yamshchikov, I. Generalization potential of large language models. *Neural Comput. Appl.* **37**, 1973–1997 (2025).
12. Kumar, A., Makhija, P. & Gupta, A. Noisy text data: Achilles' heel of bert. In *6th Workshop on Noisy User-Generated Text (W-NUT 2020)*, 16–21 (2020).
13. Zhang, L., Tian, Z.-H., Zhou, W. & Wang, W. Learning from long-tailed noisy data with sample selection and balanced loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Main Track, IJCAI-24*, 5471–5480 (2024).
14. Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint* arXiv:1909.11942 (2019).
15. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint* arXiv:1910.01108 (2019).
16. Gamage, G., De Silva, D. & Mills, N. E. A. Emotion aware: An artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis. *J Big Data* **11**, 93. https://doi.org/10.1186/s40537-024-00953-2 (2024).
17. Maity, K., Saha, S. & Bhattacharyya, P. Emoji, sentiment and emotion aided cyberbullying detection in hinglish. *IEEE Trans. Comput. Soc. Syst.* **10**, 2411–2420 (2023).
18. Bashir, M. F. et al. Context-aware emotion detection from low-resource urdu language using deep neural network. *ACM Trans. Inf. Syst.* **22**, 1–30 (2023).
19. Kitanovski, A., Toshevska, M. & Mirceva, G. Distilbert and roberta models for identification of fake news. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 1102–1106, https://doi.org/10.23919/MIPRO57284.2023.10159740(2023).
20. Chen, Y. & Ye, C. Satirical news headline detection based on bert+lstm. In *2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 962–966 (2024).
21. Choudhry, A., Khatri, I., Jain, M. & Vishwakarma, D. K. An emotion-aware multitask approach to fake news and rumor detection using transfer learning. *IEEE Trans. Comput. Soc. Syst.* **11**, 588–599 (2024).
22. Lai, Y. & Zhang, L. Government affairs message text classification based on roberta and textcnn. In *2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE)*, 258–262 (2023).
23. Santhiya, S. *et al.* A comparative exploration in text classification for hate speech and offensive language detection using bert-based and glove embeddings. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (2024).
24. Mladenovic, D. et al. Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers. *Sci. Rep.* **14**, 25731 (2024).
25. Wang, X. *et al.* Sentiment classification based on roberta and data augmentation. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, 260–264 (2023).
26. Saha, T., Upadhyaya, A., Saha, S. & Bhattacharyya, P. A multitask multimodal ensemble model for sentiment- and emotion-aided tweet act classification. *IEEE Trans. Comput. Soc. Syst.* **9**, 508–517 (2022).
27. Plaza-del Arco, F. M., Curry, A., Curry, A. C. & Hovy, D. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5696–5710 (Torino, Italia, 2024).
28. Acheampong, F., Mensah, H. & Chen, W. Transformer models for text-based emotion detection: A review of bert-based approaches. *Artif. Intell. Rev.* **54**, 5789–5829 (2021).
29. Mei, L., Liu, S., Wang, Y., Bi, B. & Cheng, X. Slang: New concept comprehension of large language models. *arXiv preprint* arXiv:2401.12585 (2024).
30. Cortiz, D. Exploring transformers models for emotion recognition: a comparision of bert, distilbert, roberta, xlnet and electra. In *Proceedings of the 3rd International Conference on Control, Robotics and Intelligent System*, 230–234 (Association for Computing Machinery, 2022).
31. Yan, J., Pu, P. & Jiang, L. Emotion-rgc net: A novel approach for emotion recognition in social media using roberta and graph neural networks. *PLoS ONE* **20**(3), e0318524 (2025).
32. Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H. & Smith, N. A. Probing across time: What does roberta know and when? *arXiv preprint* arXiv:2104.07885 (2021).
33. Conneau, A., Schwenk, H., Barrault, L. & Lecun, Y. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*, 1107–1116 (2017).
34. Kaggle. Emotion dataset for nlp. https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp (2020).
35. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. *arXiv preprint* arXiv:1910.03771https://doi.org/10.48550/arXiv.1910.03771(2019).
36. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
37. CardiffNLP. Twitter roberta-based model for emotion classification. https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion (n.d). Accessed: YYYY-MM-DD.
38. Delgado, R. & Tibau, X. Why Cohen's kappa should be avoided as performance measure in classification. *PLoS ONE* **14**(9), e0222916. https://doi.org/10.1371/journal.pone.0222916 (2019).
39. Li, X. *et al.* Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics.* Vol. 54, 465-476 (2020).
40. Dietterich, T. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
41. Micikevicius, P. *et al.* Mixed precision training. *arXiv preprint* arXiv:1710.03740 (2017).

42. Kumar, A. Sentiment and emotions of tweets dataset. https://www.kaggle.com/datasets/ankitkumar2635/sentiment-and-emotions-of-tweets (2021).
43. Sharma, K. Sentiment and emotion analysis dataset. https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset (2024).
44. Ameer, I. et al. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.* **213**, 118534 (2023).
45. Helal, N. et al. A contextual-based approach for sarcasm detection. *Sci. Rep.* **14**, 15415 (2024).
46. Liu, H., Yin, Q. & Wang, W. Y. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (eds Korhonen, A. et al.) 5570–5581 https://doi.org/10.18653/v1/P19-1560 (Association for Computational Linguistics, 2019).

Acknowledgements

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU250697].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.