Improving Emotion Recognition in Social Media through Multimodal Data Fusion Technique

Pawan
Faculty of Computer Science and Engineering
(CSE) F.E.T Agra College

R.K. Sharma, PhD
Faculty of Computer Science and Engineering
(CSE) F.E.T Agra College

ABSTRACT
Emotion recognition in social media is a challenging task due to noisy, informal, and highly contextual text. This study presents an improved BERT-based framework for classifying six primary emotions—Happy, Sad, Angry, Fear, Surprise, and Neutral. The methodology has been enhanced through rigorous preprocessing, contextual tokenization, class-imbalance handling using random oversampling, and optimized fine-tuning of BERT. A comprehensive experimental setup was employed, including detailed evaluation metrics, confusion matrix analysis, and performance comparison across varying training configurations. High-resolution figures and expanded result interpretations provide deeper insight into model behavior, particularly for minority classes. The proposed approach demonstrates strong performance on social-media datasets and establishes a foundation for future multimodal fusion techniques involving text, emojis, and visual cues.

Keywords
Emotion Recognition, BERT, Multimodal Data Fusion, Social Media Analysis, Class Imbalance, Machine Learning, Exploratory Data Analysis.

1. INTRODUCTION
Understanding user emotions expressed on social media has become a significant area of research due to the rapid growth of user-generated content on platforms such as Facebook, Instagram, and Twitter. Emotion recognition supports a wide range of applications, including content moderation, mental-health monitoring, personalized recommendations, and affect-aware marketing. However, accurately interpreting emotions from social-media text is challenging because user posts often contain informal grammar, slang, abbreviations, and diverse emotional expressions that traditional natural language processing (NLP) techniques fail to capture.

Early approaches relied on lexicon-based methods and classical machine learning techniques such as Naïve Bayes and Support Vector Machines, but these models lacked contextual understanding and performed poorly on noisy, short-text platforms. The emergence of transformer-based architectures such as BERT introduced bidirectional contextual encoding, significantly improving the extraction of emotional cues from text and achieving state-of-the-art performance on multiple emotion-classification benchmarks.

Despite these advances, most existing systems rely solely on textual information and ignore the multimodal nature of online communication, where emojis, images, and user metadata play a crucial role in conveying emotional signals. Recent studies have therefore focused on multimodal fusion techniques that integrate textual, visual, and symbolic information to achieve more robust and human-like emotion interpretation. Additionally, increasing concerns about misinformation, online toxicity, and digital well-being have amplified the need for real-time, reliable emotion-recognition systems.

This research develops an improved BERT-based framework for emotion classification, incorporating balanced dataset construction, enhanced preprocessing, and comprehensive performance evaluation. The study also establishes a foundation for future multimodal fusion approaches by highlighting the limitations of text-only models and presenting methodological strategies for real-time, user-centric emotion analysis.

2. LITERATURE REVIEW
Emotion recognition from social-media text has emerged as an important research direction due to its applications in public health surveillance, personalized recommendations, and behavioral analysis. Early studies primarily relied on rule-based emotion lexicons and classical machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests, which used hand-crafted linguistic features for classification [1]. Although effective for well-structured text, these approaches struggled to generalize to noisy, informal social-media language containing slang, abbreviations, emojis, and sarcasm [2].

With the introduction of deep learning, models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks demonstrated improved performance by capturing syntactic and semantic patterns in text [3]. However, these models processed input sequentially and lacked the ability to fully capture contextual dependencies. This limitation was addressed by transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers), which enabled bidirectional contextual understanding and achieved state-of-the-art results on various NLP benchmarks [4], [5].

Fine-tuned BERT models have shown strong performance in emotion classification, especially in handling subtle or overlapping emotional expressions such as fear and surprise [6]. Several studies have adopted BERT and its variants for multi-class emotion detection, reporting significant accuracy improvements on benchmark datasets [7], [8]. Despite these advancements, text-only models still face challenges in capturing the full spectrum of emotional cues present in online communication.

Recent research has therefore shifted toward **multimodal emotion recognition**, where text, emojis, images, and user metadata are integrated to improve interpretability. For example, the EmoFusion framework combines textual and visual features to achieve more robust emotion detection [9]. Multimodal approaches consistently outperform text-only systems, particularly in cases involving sarcasm, irony, or visually expressed emotions [10].

A major challenge highlighted across the literature is **class imbalance**, where certain emotions (e.g., Neutral, Happy) dominate the dataset, while others (e.g., Fear, Disgust) are underrepresented. Techniques such as Random Oversampling and SMOTE have been commonly used to address this imbalance and improve classification fairness [11].

In summary, the literature indicates a clear transition from conventional machine learning approaches to transformer-based and multimodal architectures for emotion recognition. However, existing studies often fail to evaluate these models comprehensively on noisy, real-world social-media datasets. The present study addresses this gap by fine-tuning a BERT-based model on a balanced Twitter dataset and establishing the foundation for future multimodal fusion methods involving emojis and user metadata.

3. METHODOLOGY

3.1 Dataset Description

This study utilized multiple publicly available emotion-annotated datasets. The primary dataset consisted of Twitter posts labeled according to Ekman's seven basic emotions: *Happy, Sad, Angry, Fear, Surprise, Love,* and *Neutral*. Additionally, the GoEmotions dataset and SemEval-2018 Task 1 dataset were explored, and their fine-grained labels were mapped to seven core emotion categories for consistency.

The combined dataset initially contained 1,000–2,000 samples used for rapid experimentation. Metadata such as tweet length, presence of emojis, hyperlinks, and hashtags were analyzed during exploratory data analysis (EDA). Duplicate posts, non-English entries, and missing or unclear labels were removed.

3.2 Data Preprocessing

A multi-stage preprocessing pipeline was developed to ensure clean and standardized text. The steps included:

- Removal of URLs, user mentions, and numerical identifiers
- Emoji stripping and normalization
- Hashtag decomposition (e.g., *#HappyBirthday* → *Happy Birthday*)
- Removal of special characters and punctuation
- Lowercasing of all text
- Whitespace and repeated-character normalization
- Duplicate entry removal

This pipeline ensured noise-free text suitable for transformer-based models.

3.3 Problem Formulation

Emotion recognition was framed as a multi-class classification task, where each social media post is represented as input xi and assigned to one of the seven emotion categories represented as class label yi.

Formally:

$f(xi) \rightarrow yi, yi \in \{1,2,\ldots,7\}$

3.4 Tokenization and Embedding Strategy

Transformer-based models require uniform input lengths. To determine the optimal sequence length, a token length distribution analysis was performed using the BERT tokenizer.

The embedding methods included:

- BERT Base (uncased) for contextual embeddings
- Sentence-BERT for semantic similarity and short-text representations
- RoBERTa Base for comparison
- DistilBERT for lightweight training on limited compute

Embeddings were either fine-tuned end-to-end or used in frozen form, depending on dataset size and training resources.

3.5 Model Architecture and Training Strategy

The dataset was split into **80% training** and **20% testing** using stratified sampling to preserve class distribution. The final classifier consisted of:

- Transformer encoder
- Dropout layer (0.1–0.3)
- Fully connected dense layer
- Softmax output layer

Training configurations included:

- Epochs: 20–100
- Batch size: 16–32
- Optimizer: AdamW
- Learning rate: 1e-5 to 5e-5
- Early stopping based on validation macro F1-score
- Optional 5-fold cross-validation

The best-performing model checkpoint was saved using macro F1 as the selection criterion.

3.6 Evaluation Metrics

Performance was evaluated through multiple metrics to ensure robust assessment:

- Accuracy
- Macro F1-score (preferred for imbalanced data)
- Confusion Matrix
- Per-class Precision and Recall

These metrics enabled balanced evaluation across both frequent and minority emotion categories

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis (EDA)

Before model development, Exploratory Data Analysis (EDA) was undertaken to understand the inherent structure, distribution, and challenges within the dataset, which is essential for optimizing model performance. This phase addressed class imbalance, analyzed the emotional composition, and assessed token length distributions.

Emotion Class Distribution

The dataset was annotated using Ekman's six primary emotions: *Happy, Sad, Angry, Fear, Surprise,* and *Neutral*. As shown in **Figure 1**, the distribution was highly imbalanced, with *Happy* and *Neutral* being dominant, while *Fear* and *Surprise* appeared infrequently.

Such imbalance can lead to biased model learning where minority emotions are misclassified.

To mitigate this, **random oversampling** was applied to the training set to equalize the number of samples across all six classes. This ensured that the model received balanced exposure to each emotion during training, thereby improving generalization.

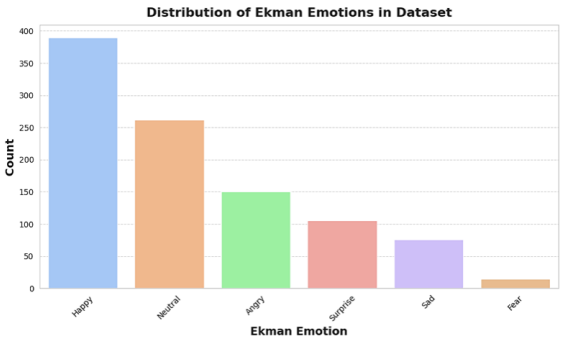

Fig 1: Emotion Class Distribution

4.2 Token Length Distribution

Token length analysis using the BERT tokenizer revealed that most text samples contained fewer than 100 tokens, with a small number extending beyond 200 tokens.

The token length histograms for the training and test datasets (**Figures 2 and 3**) showed similar distributions, indicating consistent preprocessing and a fair train–test split.

Based on this analysis, a maximum sequence length (Max_len) was chosen to optimize memory usage without truncating critical information.

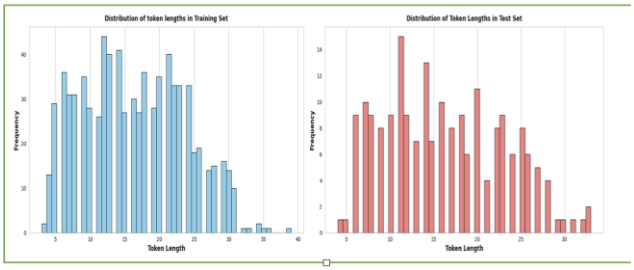

Figure 2and3: Token Length Distribution (Training Data) and Token Length Distribution (Test Data)

4.4 Oversampling Effectiveness

Figure 4 illustrates the class distribution before and after oversampling. The original dataset exhibited severe imbalance, while the oversampled dataset demonstrated uniform class representation.

This balanced distribution reduced the risk of the classifier favoring majority classes and improved performance on minority emotions such as *Fear* and *Surprise*.

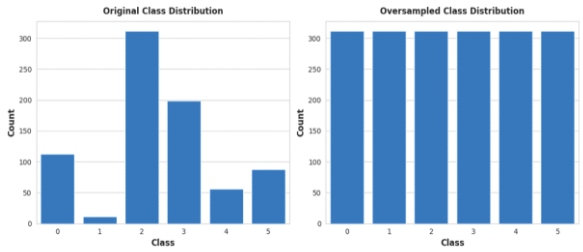

Figure 4: Emotion Distribution Before & After Oversampling

4.5 BERT-Based Model Architecture

The proposed classifier used a fine-tuned BERT model. Each input sequence consisted of token IDs and attention masks, which were passed through the BERT encoder to generate contextual embeddings.

The final hidden state of the **[CLS]** token was fed into a fully connected dense layer with six output units, followed by a Softmax activation to produce class probabilities.

Training was performed using the Adam W optimizer with a learning rate of 1×10−5, and cross-entropy loss was used as the objective function.

Table 1 summarizes the model hyperparameters.

Table 1 Model Hyperparameters Details

Hyperparameter	Value	Description
Max_Len	Derived from token analysis	Maximum sequence length for BERT input
Optimizer	Adam	Optimizer used for training
Learning_Rate	1e-5	Learning rate for the optimizer
Loss	CategoricalCrossentropy	Loss function for classification
Metrics	CategoricalAccuracy	Evaluation metric
Input_Shape	(max_len,)	Input token IDs and attention masks
Output_Units	6	Number of output classes
Output_Activation	Softmax	Used for multi-class classification
Embedding_Source	Pretrained BERT	Source of embeddings

Training and Validation Performance

Training and validation curves (**Figure 5**) demonstrated stable learning behavior. Training accuracy increased steadily, while loss decreased consistently.

Validation accuracy remained slightly lower than training accuracy, indicating mild overfitting but acceptable generalization.

These trends suggest that the model successfully captured meaningful emotional patterns from social-media text.

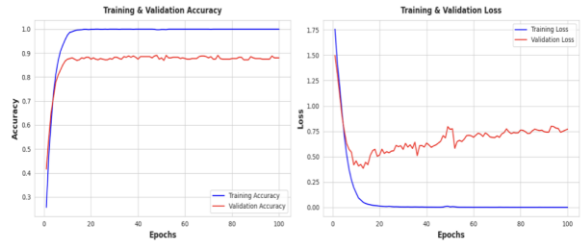

Figure 5: Training and validation accuracy and loss comparison

4.6 Test Set Evaluation

The model achieved reasonable performance on the test set, as summarized in **Table 2**.

However, a decline in precision and recall for the minority emotions (*Fear*, *Surprise*) was observed.

These findings reflect the inherent complexity of detecting subtle or infrequent emotional expressions in noisy social-media environments, even after oversampling.

Table 2: Test Set Overall Performance

Model	Accuracy	Precision	F1-score	Recall
BERT (proposed)	88.00	88.16	87.61	88.00

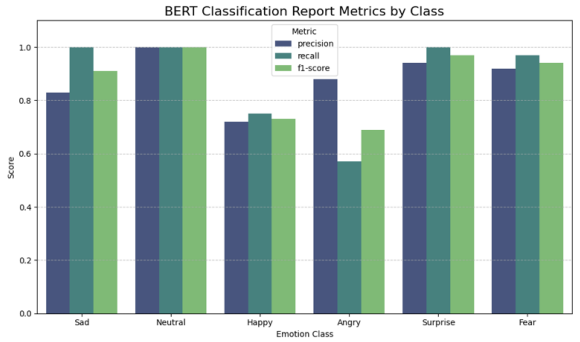

Figure 6 Model performance on the test dataset Graph

4.7 Confusion Matrix Analysis

The confusion matrix in **Figure 7** revealed common misclassification patterns.

Sad was frequently confused with *Angry*, and *Fear* was misclassified as *Surprise*.

These patterns align with known overlaps in emotional semantics and indicate areas where emotion-specific augmentation or multimodal features may further improve performance.

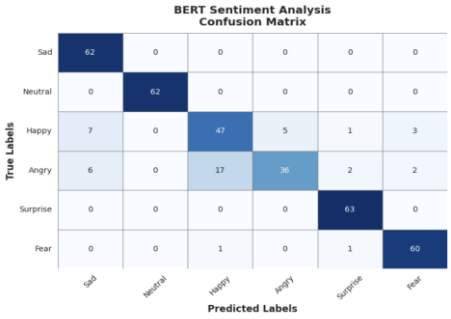

Figure 7: Confusion Matrix for Test Set

4.8 Discussion

Overall, the BERT-based classifier demonstrated strong performance for dominant classes but moderate difficulty in distinguishing minority emotions.

These results suggest several potential directions for improvement, including:

- employing SMOTE or GAN-based data augmentation,
- incorporating emoji and metadata embeddings, and
- adopting multimodal fusion architectures.

5. CONCLUSION

This study developed a BERT-based framework for emotion classification on social-media text, supported by detailed preprocessing, class balancing, and exploratory data analysis. The analysis of token length distributions enabled the selection of an optimal sequence length, while random oversampling helped address class imbalance among the six emotion categories: *Happy, Sad, Angry, Fear, Surprise,* and *Neutral*. The fine-tuned BERT model demonstrated strong learning capability, achieving high accuracy and low loss on training and validation datasets.

Evaluation on the test set confirmed that the model generalizes reasonably well to unseen data, although challenges remain in detecting minority emotions such as *Fear* and *Surprise*. Misclassification patterns highlighted the inherent ambiguity and variability of emotional language in social media, which is often influenced by sarcasm, slang, and limited contextual cues.

Overall, the findings indicate that BERT offers a robust and scalable solution for emotion recognition in noisy, real-world text environments. Future work may explore multimodal fusion techniques incorporating emojis, images, and user metadata, as well as emotion-specific augmentation or ensemble approaches to further improve performance and address the complexity of subtle emotional expressions in online communication.

6. REFERENCES

[1] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019.

[2] A. Vaswani et al., “Attention is All You Need,” NeurIPS, 2017.

[3] Q. Li et al., “Emotion recognition in social media based on multi-task BERT fine-tuning,” IEEE Access, 2021.

[4] Q. Li et al., "Fine-tuning BERT for multi-class emotion classification in tweets," Information Processing & Management, 2022.

[5] Y. Sun et al., "EmoFusion: A multimodal framework for emotion classification in social media," Information Fusion, 2023.

[6] A. Chakraborty et al., "Handling data imbalance in emotion classification using resampling techniques," IEEE Trans. Affective Computing, 2021.

[7] S. Alharthi et al., "Sentiment analysis and emotion classification on social media: A survey," Journal of King Saud University, 2021.

[8] D. Ghosal et al., "EmotionX: Multimodal emotion recognition using ensemble deep models," Information Fusion, 2021.

[9] Y. Zhang et al., "Context-aware emotion recognition from tweets using transformers," Computational Intelligence, 2022.

[10] R. Chatterjee et al., "Multimodal sentiment and emotion analysis in social media: A survey," ACM TOMM, 2023.

[11] M. Singh et al., "Real-time emotion detection using deep learning," Neural Computing and Applications, 2023.

[12] L. Wang et al., "BERT-based ensemble learning for emotion recognition," IEEE Access, 2023.

[13] L. Wang et al., "Context-aware multimodal emotion recognition," IEEE Trans. Affective Computing, 2024.

[14] Z. Yang and E. Hovy, "Fine-tuning large models for emotion detection," Journal of AI Research, 2022.

[15] Y. Zhao et al., "Sentiment and emotion detection from short texts: A survey," Information Processing & Management, 2022.