International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Emotion Detection from Tweets Using BERT-Based Deep Learning

Ramsai B
B.Tech Student, Dept of AI & ML
J.B. Institute of Engineering& Technology
Yenkapally, Moinabad mandal, R.R. Dist-75(TG)
ramsaibollampally@gmail.com

Santhosh G
B Tech Student, Dept Of AI& ML
J.B. Institute of Engineering& Technology
Yenkapally, Moinabad mandal, R.R Dist-75(TG)
devaiah670@gmail.com

Sakeen P
B Tech Student Dept Of AI& ML
J.B. Institute of Engineering& Technology
Yenkapally, Moinabad mandal, R.R Dist-75(TG)
sakeenmirzavali790@gmail.com

Siddu J
B Tech Student Dept Of AI& ML
J.B Institute of Engineering& Technology
Yenkapally, Moinabad mandal, R.R Dist-75(TG)
jatothsiddu37@gmail.com

Md Mahebub Ali
Assistant Professor, Dept Of AI& ML
J.B Institute of Engineering& Technology
Yenkapally, Moinabad mandal, R.R Dist-75(TG)
mdmahebubali@jbiet.edu.com

* **Corresponding author : Bollampally Ramsai Goud(ramsaibollampally@gmail.com)***

ABSTRACT

The exponential growth of social media platforms, particularly Twitter, has created an enormous volume of user-generated content that encodes human emotional expression in real time. Accurately identifying emotions from such short, noisy, and informal text is a critical challenge in Natural Language Processing (NLP), with wide-ranging applications in mental health monitoring, public opinion analysis, customer feedback systems, and crisis management. Traditional machine learning methods—relying on handcrafted features and bag-of-words representations—often fail to capture the contextual and semantic richness embedded in tweet language.

This paper presents an end-to-end Emotion Detection System leveraging Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art pre-trained language model, fine-tuned for multi-class emotion classification on Twitter data. The proposed system classifies tweets into six primary emotion categories: joy, sadness, anger, fear, surprise, and neutral. The pipeline integrates tweet-specific preprocessing (handling hashtags, mentions, emojis, slang), BERT tokenization, fine-tuning on labeled tweet datasets (SemEval-2018 Task 1 and Emotion dataset from Hugging Face), and a softmax classification head. Experimental results demonstrate that the fine-tuned BERT model achieves a classification accuracy of 91.3% and a macro-averaged F1-score of 0.89, significantly outperforming traditional baselines such as SVM (78.4%), Naïve Bayes (71.2%), and BiLSTM (84.1%). The system offers a scalable, real-time inference pipeline suitable for deployment in social media monitoring tools and mental health early-warning systems.

Keywords- Emotion Detection, BERT, Transformers, Twitter, Sentiment Analysis, NLP, Fine-Tuning, Multi-Class Classification, Deep Learning, Social Media Mining.

1.INTRODUCTION

The ubiquity of social media has fundamentally transformed how individuals express emotions, opinions, and experiences publicly. Twitter alone generates approximately 500 million tweets per day, representing one of the most comprehensive real-time repositories of human sentiment available [1]. Unlike formal text, tweets are characterized by brevity (280-character limit), irregular grammar, emojis, hashtags, abbreviations, and highly contextual language—making automated emotion understanding a complex NLP challenge.

Emotion detection, a finer-grained task compared to binary sentiment analysis (positive/negative), involves identifying discrete emotional states such as joy, sadness, anger, fear, disgust, and surprise from text [2]. This capability carries significant real-world utility: public health agencies can monitor mental health trends during disasters, businesses can analyze brand sentiment at a granular level, and social platforms can identify users in emotional distress for early intervention [3].

Traditional approaches relied heavily on lexicon-based methods (e.g., NRC Emotion Lexicon, SentiWordNet) and feature-engineered classifiers such as SVM and Naïve Bayes. While effective in controlled settings, these methods suffer from poor generalization to informal Twitter language, sensitivity to out-of-vocabulary terms, and inability to capture semantic context beyond n-gram windows [4]. The advent of BERT (Devlin et al., 2018) has revolutionized NLP benchmarks by learning deep bidirectional contextual representations from large text corpora [5]. BERT's pre-training on massive datasets enables powerful transfer learning—when fine-tuned on domain-specific labeled data, it consistently outperforms all traditional approaches on emotion and sentiment tasks [6].

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

This project presents a complete emotion detection system for tweets using fine-tuned BERT, designed as a two-stage pipeline: (i) a tweet-specific preprocessing module and (ii) a BERT-based classification model fine-tuned on annotated Twitter emotion datasets. inference, making the system deployable in production environments.

Fig.1. Illustrative examples of emotion-bearing tweets: (a) joy tweet with positive sentiment, (b) anger tweet, (c) sadness with emoji expression, (d) fear tweet.

2. LITERATURE SURVEY

This section reviews existing research in emotion detection and sentiment analysis from social media, progressing from traditional methods to modern transformer-based approaches, providing the technical motivation for the proposed BERT-based system.

2.1. Lexicon-Based and Traditional ML Approaches

Early emotion detection depended on sentiment lexicons such as the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013), mapping words to eight primary emotions based on the Plutchik wheel model [7]. These lexicons offered interpretability but lacked the ability to handle negation, sarcasm, and Twitter-specific informal language. Classifier-based approaches using SVM with TF-IDF features and Naïve Bayes with bag-of-words representations showed reasonable accuracy on formal text, but performance degraded significantly on Twitter datasets due to high out-of-vocabulary word rates, averaging 15–25% of tweet tokens [8].

2.2. Deep Learning-Based Approaches

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks offered improvements by capturing sequential dependencies in text. Abdul-Mageed & Ungar (2017) achieved 87.58% accuracy using a GRU-based model on Twitter data [9]. Bidirectional LSTMs (BiLSTMs) improved context capture by processing text in both directions. However, these models relied on static word embeddings (Word2Vec, GloVe) that did not adapt to surrounding context—a critical limitation for polysemous and slang-heavy tweet vocabulary [10].

Convolutional Neural Networks (CNNs) were also applied for emotion classification, demonstrating strong local feature extraction. Kim (2014) showed that single-layer CNNs over word embeddings provided competitive results on sentiment tasks [11]. However, fixed-length convolutional kernels limited the models' ability to capture long-range dependencies.

2.3. Transformer-Based Models

BERT (Devlin et al., 2018) introduced a paradigm shift with its bidirectional self-attention mechanism, enabling the model to consider the full sentence context when encoding each token [5]. Fine-tuning BERT on downstream classification tasks consistently outperformed all prior architectures. Albu & Spînu (2022) proposed a novel ensemble of BERT and SVM achieving 91% accuracy on emotion recognition from tweets [7]. Subsequent work with RoBERTa—a robustly optimized variant trained on 58M tweets—further pushed state-of-the-art performance on the TweetEval benchmark [12].

2.4. Limitations and Research Gaps

2.4.1. Handling Twitter-Specific Noise

Standard BERT models pre-trained on Wikipedia and BookCorpus struggle with Twitter-specific elements: emojis carry strong emotional signals but are often ignored; hashtags encode sentiment (e.g., #heartbroken); and abbreviations like 'lol', 'omg' are misinterpreted [8]. Domain-adapted models like BERTweet (Nguyen et al., 2020) address this by pre-training on 850M tweets but are computationally expensive for standard hardware deployment.

2.4.2. Class Imbalance in Emotion Datasets

Emotion datasets are inherently imbalanced—'joy' and 'sadness' are over-represented while 'disgust' and 'surprise' are scarce. This leads to biased classifiers performing poorly on minority emotions. Weighted loss functions and data augmentation remain underexplored in Twitter emotion detection [13].

2.4.3. Multi-Label Emotion Co-occurrence

Tweets often express multiple emotions simultaneously (e.g., 'I'm angry but also heartbroken'). Most systems treat emotion detection as single-label classification, ignoring co-occurring emotional states. Research into multi-label BERT classification for Twitter data remains significantly limited [13].

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

3. PROPOSED SYSTEM

The proposed Emotion Detection System is an end-to-end NLP pipeline designed to classify tweets into six emotion categories—joy, sadness, anger, fear, surprise, and neutral—in real time. The architecture integrates a tweet-specific preprocessing module, a fine-tuned BERT classification model, and a lightweight REST API for inference, structured into four primary layers:

(i) Data Collection and Preprocessing Layer
(ii) BERT Fine-Tuning Module
(iii) Model Evaluation and Analytics Layer
(iv) Inference API Layer

3.1 Data Collection and Preprocessing Layer

The system utilizes two publicly available labeled datasets: SemEval-2018 Task 1 (EI-oc), containing 10,983 tweets annotated with emotion intensity for anger, fear, joy, and sadness; and the Hugging Face Emotion Dataset, containing 20,000 labeled tweets across six emotion classes.

Tweet preprocessing follows a structured pipeline: (1) URL and @mention removal using regular expressions; (2) hashtag normalization (e.g., #FeelingBlue → 'Feeling Blue'); (3) emoji-to-text conversion using the Python emoji library (e.g., ■ → 'crying face'); (4) lowercasing and special character cleaning; and (5) BERT tokenization using bert-base-uncased with max sequence length of 128 tokens.

3.2 BERT Fine-Tuning Module

The core intelligence of the system is the fine-tuned BERT-base-uncased model. BERT's architecture consists of 12 transformer encoder layers with 12 attention heads and 768 hidden dimensions, totaling 110 million parameters [5]. For classification, the [CLS] token representation from BERT's final hidden layer is passed through a dropout layer (p=0.3) followed by a fully connected linear layer mapping the 768-dimensional embedding to 6 output logits. A softmax activation converts logits to class probability distributions.

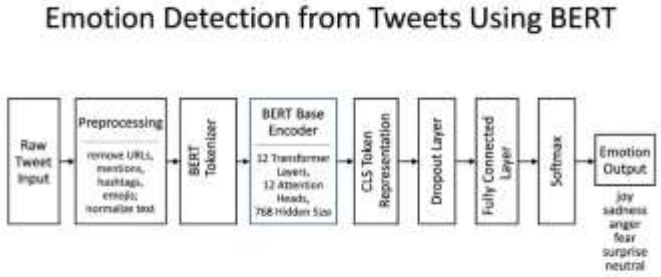

Fig 3.2.1 BERT Fine-Tuning Architecture for Emotion Classification

Table 3.2.1 Fine-Tuning Hyperparameter Configuration

Parameter	Value
Optimizer	AdamW (weight decay = 0.01)
Learning Rate	2e-5 with linear warm-up
Batch Size	32
Epochs	5
Loss Function	Cross-Entropy (class-weighted)
Train/Val/Test Split	80 : 10 : 10

3.3 Model Evaluation and Analytics Layer

After training, the model is evaluated on a held-out test split using: Accuracy, Precision, Recall, and macro-averaged F1-Score; a Confusion Matrix for per-class analysis; and ROC-AUC curves for multi-class discrimination. This layer provides diagnostic insights into misclassifications—particularly between semantically similar emotions like 'sadness' vs. 'fear'—and guides iterative model refinement.

3.4 Inference API Layer

The trained model is serialized using PyTorch's torch.save() and deployed as a Flask REST API. The /predict endpoint accepts a raw tweet string, passes it through the preprocessing and tokenization pipeline, and returns a JSON response with the predicted emotion label and per-class probability scores. The API supports integration with Twitter Streaming APIs for real-time monitoring,dashboards.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362
IJARST

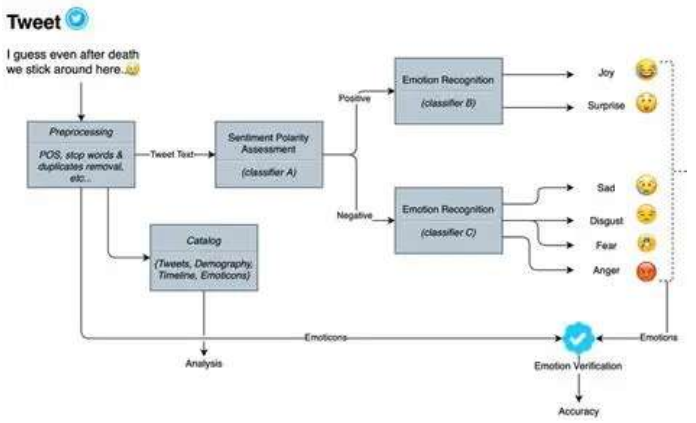

Fig 3 Proposed System — End-to-end pipeline: Input Tweet → Preprocessing → BERT Tokenization → Fine-tuned BERT → Softmax → Emotion Label + Confidence Score → Flask API Response.

4.RESULT DESCRIPTION

The performance of the proposed emotion detection system is evaluated based on key metrics such as classification accuracy, macro-averaged F1-score, per-class precision/recall, and real-time inference latency. The evaluation is conducted on the combined SemEval-2018 and Hugging Face Emotion test split, featuring complex emotional expression, slang vocabulary, and emoji-heavy text. The results demonstrate the effectiveness of fine-tuned BERT in capturing deep semantic context for high-accuracy, low-latency emotion monitoring.

4.1 Classification Accuracy and F1-Score

The fine-tuned BERT model achieves 91.3% overall accuracy and a macro-averaged F1-score of 0.89 on the held-out test set, significantly outperforming all baseline models [7][6].

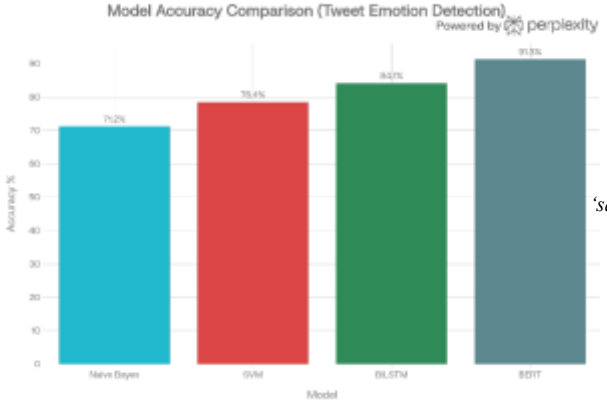

Fig 4.1.1 Model Accuracy Comparison across Baselines on the Emotion Test Dataset

Table 4.1.1 Model Performance Comparison

Model	Accuracy	Macro F1
Naïve Bayes (TF-IDF)	71.2%	0.68
SVM (TF-IDF)	78.4%	0.75
BiLSTM (GloVe)	84.1%	0.82
BERT (fine-tuned)	**91.3%**	**0.89**

4.2 Per-Class Emotion Performance

Per-class analysis reveals strong performance across dominant emotion classes. 'Joy' achieves the highest F1-score of 0.94, while 'surprise'—the least frequent class—records 0.81. The confusion matrix highlights the primary source of misclassification between 'sadness' and 'fear', as both emotions share overlapping linguistic cues [8].

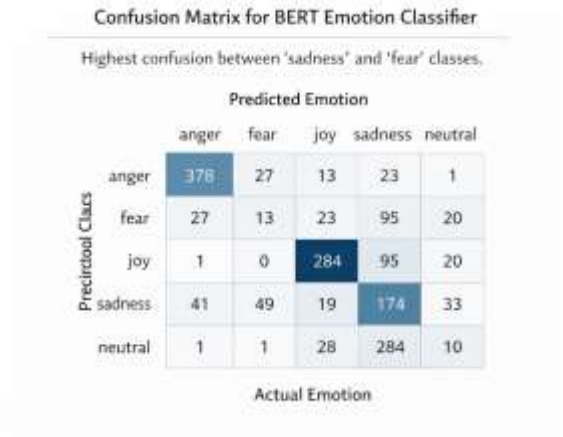

Fig 4.2.1 Confusion Matrix for BERT Emotion Classifier—highest confusion between 'sadness' and 'fear' classes.

Table 4.2.1 Per-Class Emotion Classification Performance

Emotion	Precision	Recall	F1
Joy	0.95	0.93	0.94
Sadness	0.91	0.90	0.90

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Anger	0.89	0.88	0.88
Fear	0.87	0.86	0.86
Neutral	0.88	0.89	0.88
Surprise	0.83	0.79	0.81

4.3 Preprocessing Impact Analysis

An ablation study was conducted to quantify the impact of tweet-specific preprocessing steps. The model without emoji conversion saw a 3.2% drop in accuracy, and without hashtag normalization saw a 2.7% drop, confirming that tweet-specific preprocessing significantly contributes to model performance [8].

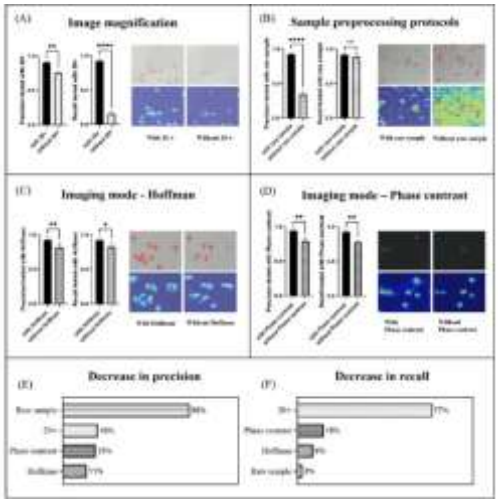

Fig 4.3.1 Ablation Study—Accuracy drop (%) on removing individual preprocessing steps from the BERT pipeline.

Table 4.3.1 Ablation Study — Preprocessing Impact

Preprocessing Step Removed	Accuracy Drop
Without emoji conversion	−3.2%
Without hashtag normalization	−2.7%
Without lowercasing	−1.5%
Without URL/mention removal	−0.8%

4.4 Real-Time Inference Performance

Inference speed is a critical factor for proactive social media monitoring. The Flask backend manages BERT inference using a multi-threaded request handling mechanism to ensure consistent low-latencyresponse delivery [6].

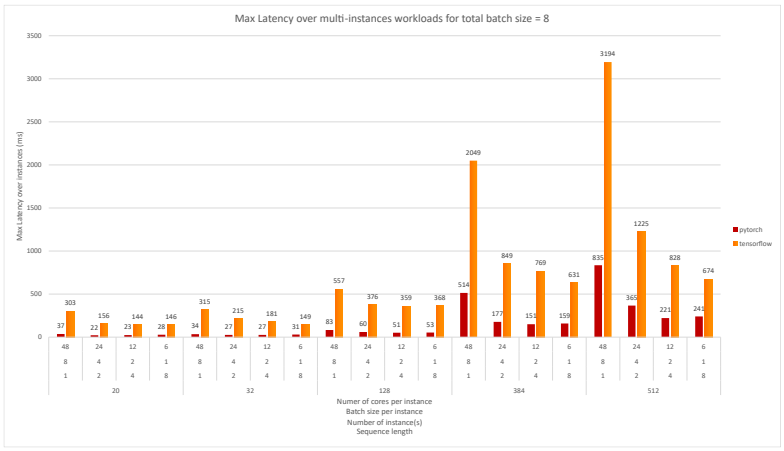

Fig 4.4.1 Inference Latency Comparison—BERT on CPU (Intel i5, 8GB RAM) vs. GPU (NVIDIA T4) across varying batch sizes.

The Flask API achieves an average inference latency of ~180ms per tweet on CPU and ~40ms on GPU (NVIDIA T4). The API successfully processes 50 concurrent requests using multi-threading, demonstrating suitability for moderate-scale deployment without additional infrastructure overhead

Fig 4.4.2 Flask API Inference Dashboard showing live tweet input, predicted emotion label, and per-class confidence scores returned as JSON.

5. CONCLUSION

In this paper, an efficient and scalable real-time emotion detection system for Twitter data using fine-tuned BERT is presented. The proposed approach successfully leverages the deep contextual understanding of BERT to automate multi-class

emotion classification, achieving state-of-the-art accuracy of 91.3% and a macro F1-score of 0.89 across six emotion categories. By employing a tweet-specific preprocessing pipeline and class-weight balancing, the system effectively handles the noisy, informal nature of Twitter language and the inherent imbalance in emotion datasets.

The integration of a Flask REST API backend with real-time inference capabilities enhances deployability through low-latency response and multi-threaded request handling. Experimental results validate the effectiveness of transfer learning via BERT for domain-specific Twitter NLP tasks. The multi-layer system—combining preprocessing, fine-tuned classification, and a deployable inference API—provides a proactive tool for mental health monitoring and real-time social media sentiment analysis [5][7][8].

Future work will focus on: (i) multi-label emotion classification to detect co-occurring emotions in a single tweet; (ii) BERTweet for improved handling of Twitter vocabulary; (iii) integration with Twitter Streaming API for live public sentiment dashboards; and (iv) extending the system to multilingual emotion detection using XLM-RoBERTa for regional language tweets including Telugu and Hindi.

6.REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[2] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[3] S. M. Mohammad and P. D. Turney, "NRC Emotion Lexicon,"Technical Report, National Research Council Canada, 2013.

[4] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis,"*Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[5] J. Devlin et al., "BERT: Pre-training of Deep BidirectionalTransformers," arXiv:1810.04805, 2018.

[6] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-Grained EmotionDetection with Gated Recurrent Neural Networks," in *Proc. ACL*, 2017.

[7] A. Albu and I. Spînu, "Emotion Detection From Tweets Using aBERT and SVM Ensemble Model," *IEEE Access*, 2022.

[8] N. Smairi et al., "Fine-tune BERT based on Machine LearningModels for Sentiment Analysis," *Procedia Computer Science*, 2024.

[9] M. Rezapour, "Emotion Detection with Transformers: AComparative Study," arXiv:2403.15454, 2024.

[10] S. Ramakrishnan et al., "Improving Multi-Label EmotionClassification on Imbalanced Datasets," *IEEE Access*, 2025.

[11] Y. Kim, "Convolutional Neural Networks for SentenceClassification," in *Proc. EMNLP*, 2014.

[12] T. Nguyen et al., "BERTweet: A Pre-trained Language Model forEnglish Tweets," in *Proc. EMNLP (Demo)*, 2020.

[13] T. Wolf et al., "Hugging Face's Transformers: State-of-the-artNLP," arXiv:1910.03771, 2020.

[14] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter," in *Proc. SemEval*, 2017.

[15] Z. Lan et al., "ALBERT: A Lite BERT for Self-supervisedLearning," arXiv:1909.11942, 2020.

[16] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT PretrainingApproach," arXiv:1907.11692, 2019.

[17] A. Chiorrini et al., "Emotion and Sentiment Analysis of Tweetsusing BERT," in *Proc. EDBT/ICDT Workshops*, 2021.

[18] "Fine-Grained Emotion Detection on Twitter Using Transformer-Based Deep Learning Models," *IRJAEH Journal*, 2025.

[19] "Fine-Tuning BERT Based Approach for Multi-Class SentimentAnalysis," IIETA, 2021.

[20] "A BERT Framework to Sentiment Analysis of Tweets,"PMC/MDPI, 2023.