Emotional Context Detection for Content Moderation

Mr. Piyush. R. Kulkarni
Assistant Professor, Computer Engineering , Guru Gobind Singh College of Engineering and Research Centre, Nashik (GCOERC)

Mr. Sarang Arvind Yerne, Mr. Saurabh Milind Sirsat, Mr. Aksay Arun Bacchav, Mr. Kunal Dhiraj Chaudhari
Students, Department of Computer enginerring, Guru Gobind Singh College of Engineering and Research centre, Nashik (GCOERC)

Abstract: The explosive proliferation of social media platforms like Twitter has generated enormous volumes of text-based data capturing users' feelings, perspectives, and attitudes. Automatically interpreting these emotional signals offers significant value across psychological research, consumer behavior analysis, and real-time public monitoring systems. This project introduces an Emotion Detection System that utilizes DistilBERT, a compact yet powerful transformer-based architecture, for high-quality feature extraction from raw textual inputs. The system initially applies preprocessing operations including tokenization and sequence padding to structure input text for the pre-trained DistilBERT model. The resulting contextual sentence representations are subsequently passed through fully connected dense layers for feature refinement, followed by a fine-tuning mechanism to further boost classification accuracy. Logistic Regression is then applied to categorize text into distinct emotional classes such as happiness, anger, sadness, fear, and surprise. This combined methodology bridges deep contextual language understanding of transformer-based embeddings with the transparency and efficiency of conventional machine learning techniques. The proposed framework delivers effective emotion recognition while maintaining low computational overhead, making it well-suited for real-time social media analytics and sentiment-driven applications within integrated campus management infrastructure.

Keywords: - Emotion Detection, DistilBERT, Logistic Regression, Natural Language Processing (NLP), Sentiment Analysis, Feature Extraction, Text Classification, Deep Learning, Social Media Analytics.

1. INTRODUCTION

The rapid evolution of internet technology has contributed to a substantial surge in multimedia content, encompassing text, audio, images, and video. Among these formats, written text remains one of the most extensively circulated forms of digital information online. Social media platforms produce millions of posts daily, creating a pressing demand for effective processing of unstructured textual content. This has sparked growing interest in opinion mining and emotional analysis. Users openly share their thoughts, perspectives, and feelings on digital platforms, generating rich data for computational analysis. Numerous organizations capitalize on this user-generated content to strengthen internal decision-making processes and better understand prevailing public attitudes.

Emotion detection (ED) involves identifying particular emotional states such as joy, sadness, disappointment, and fear from available data. Emotions can be recognized through multiple channels, including vocal tone, facial cues, physical gestures, and written language. The underlying assumption is that an individual's emotional condition shapes their choice of words. Emotion recognition carries practical significance across several domains, including healthcare, education, and marketing. In healthcare, emotional well-being is closely tied to physical health prolonged depression can deteriorate overall health, whereas positive emotional states may accelerate recovery. In educational settings, emotions directly shape learning outcomes, with favorable emotional states improving focus and cognitive engagement. In marketing, emotionally driven campaigns have gained widespread adoption due to their proven ability to attract audiences and stimulate consumer behavior.

Contemporary research has increasingly concentrated on advancing human-computer interaction by incorporating emotional intelligence derived from platforms such as Twitter, Instagram, YouTube, and Facebook. Text-based emotion detection employs Natural Language Processing (NLP) techniques to extract semantic meaning from written content. Conventional machine learning methods rely extensively on manual feature engineering, which frequently fails to capture subtle linguistic nuances. Deep learning approaches, by contrast, have demonstrated superior performance by autonomously learning intricate language patterns. Transfer learning, which applies knowledge acquired in one context to improve outcomes in another, further strengthens these frameworks. This study seeks to assess the performance of fine-tuned models in emotion classification using the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset. To maximize model performance, hyperparameter optimization of the Adam optimizer was carried out on the DistilBERT architecture. The core objective of the proposed system is to identify emotions from tweets by integrating the robust feature extraction capabilities of DistilBERT with the classification efficiency of Logistic Regression.

2. RELATED WORKS

Identifying emotions from social media text, especially content sourced from Twitter, has attracted considerable research interest owing to its relevance in sentiment analysis, public discourse monitoring, and mental health evaluation. Initial efforts in this area were predominantly built on conventional machine learning techniques, including Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression, which depended on manually crafted features such as bag-of-words representations, TF-IDF weightings, and n-gram models. While these approaches yielded acceptable results, they consistently fell short in capturing the contextual depth and semantic complexity inherent in brief, informal social media posts.

As deep learning gained momentum, architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), notably Long Short-Term Memory (LSTM) networks, were adopted for emotion and sentiment classification tasks. These models advanced the field by learning sequential linguistic patterns directly from data; however, their dependency on large annotated datasets and substantial computational resources posed practical limitations for real-world deployment.

The emergence of transformer-based language models represented a landmark shift in natural language processing capabilities. Bidirectional Encoder Representations from Transformers (BERT) delivered state-of-the-art results across numerous text classification benchmarks, including emotion and sentiment detection, by constructing deep bidirectional contextual representations of input text. Multiple studies have leveraged BERT for tweet-level emotion classification, consistently reporting accuracy gains over both traditional and recurrent architectures.

To mitigate the computational demands associated with BERT, researchers developed DistilBERT, a distilled and more lightweight variant that preserves the majority of BERT's capabilities while considerably reducing model size and processing latency. DistilBERT has proven highly effective for emotion recognition in social media contexts, particularly in scenarios requiring real-time performance or operating under resource constraints.

More recent investigations have explored hybrid frameworks that integrate transformer-based feature extraction with traditional machine learning classifiers. In these configurations, models such as BERT or DistilBERT generate dense, semantically rich embeddings, which are subsequently supplied to classifiers like Logistic Regression or SVM. Such hybrid systems combine the representational strength of transformer architectures with the interpretability, simplicity, and generalization ability of classical classifiers. Findings documented across existing literature suggest that Logistic Regression, when coupled with transformer-derived embeddings, can deliver highly competitive classification performance with comparatively lower training overhead. Drawing on these insights, the proposed system adopts a hybrid methodology that employs DistilBERT for deep contextual feature extraction and Logistic Regression for streamlined emotion classification.

3. MODULES AND DESCRIPTION

(a) Data Preprocessing

- The system accepts textual input such as tweets or user-provided text.
- Tokenization and padding are applied to prepare text for model ingestion.
- Data is cleaned and normalized to remove unwanted symbols, links, or stop-
- Words

(b) Feature Extraction Using Distil-BERT

- Input text is converted into tokenized sentences using the pre-trained Distil-BERT tokenizer.
- DistilBERT encoder transforms sentences into sentence-level feature represen-tations.
- Extracted embeddings are passed to dense layers for refined representation.

(c) Emotion Classification Using Logistic Regression

- The Logistic Regression model receives encoded features as input.
- It predicts probabilities for multiple emotion categories.
- Final output corresponds to the dominant emotion such as Happiness, Anger,Sadness, Fear, or Surprise.

(d) Fine-Tuning for Improved Accuracy

- The fine-tuning process adjusts model weights on a labeled dataset.
- Model performance is evaluated using standard metrics such as Accuracy,
- Precision, Recall, and F1-Score.
- The fine-tuned model is saved for future predictions and reusability.

4. .PROPOSED SYSTEM

The scope of this study encompasses the end-to-end pipeline of emotion detection, spanning from data acquisition through to final classification. Tweets are gathered either directly from Twitter or sourced from existing labeled datasets, and subsequently subjected to preprocessing operations aimed at eliminating noise elements such as URLs, hashtags, special characters, and stopwords. The resulting cleaned text is then channeled through the DistilBERT model, which produces dense vector representations that encode the contextual significance of each tweet. These embeddings function as high-quality input features for the downstream classification stage. A Logistic Regression classifier is then applied to these representations to assign each tweet to one of several predefined emotional categories. Model performance is assessed using widely accepted evaluation metrics including accuracy, precision, recall, and F1-score, thereby ensuring the dependability and consistency of the overall system.

The workflow originates with raw tweet acquisition, followed by a series of preprocessing procedures designed to produce clean and standardized input data. The refined text is subsequently introduced into the DistilBERT architecture, which transforms sentences into meaningful numerical encodings. These encoded representations are forwarded to the Logistic Regression classifier, which interprets the embeddings and predicts the most probable emotion label. The system's outputs are ultimately validated against held-out test data to confirm predictive accuracy and overall effectiveness. The proposed framework thus delivers a computationally efficient yet highly capable solution for real-time emotion recognition from continuously evolving social media data streams.

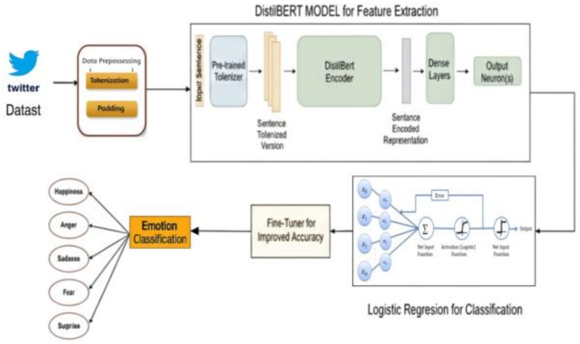

Fig 1.Model Architecture

4. LITERATURE SURVEY

Emotional Tweets Analysis on Social Media with Short Text Classification: A study targeting emotion identification in short and informal social media content, particularly Twitter posts, highlighted the inherent difficulties arising from limited context, abbreviations, colloquial language, and emoji usage. To overcome these obstacles, the researchers employed a combination of NLP techniques alongside machine learning classifiers including Support Vector Machines, Naïve Bayes, and Convolutional Neural Networks. The incorporation of semantic word embeddings alongside context-aware features enabled effective recognition of emotional states such as happiness, anger, sadness, and fear. Results confirmed that hybrid feature representations substantially improved classification accuracy, establishing a reliable framework for real-time emotion and sentiment analysis across social media environments. [1]

Hybrid Feature Extraction for Multi-Label Emotion Classification in EnglishText Messages: Another investigation proposed a hybrid feature extraction approach designed for multi-label emotion classification, acknowledging that a single message may simultaneously convey multiple emotional states. The framework combined statistical representations using TF-IDF, semantic embeddings through Word2Vec and GloVe, and deep contextual encodings via BERT and Bi-LSTM architectures. Evaluation on standard benchmark datasets revealed strong performance in identifying overlapping emotional categories. The study demonstrated that merging handcrafted linguistic features with deep learning embeddings effectively captures both syntactic structure and contextual meaning, improving detection of subtle and co-existing emotions. [2]

A Review and Critical Analysis of Multimodal Datasets for Emotional AI A comprehensive review of multimodal emotion recognition datasets examined how varied data sources — including text, speech, and facial expressions — collectively contribute to the advancement of Emotional AI. Widely adopted datasets such as IEMOCAP, MELD, and AffectNet were analyzed with respect to their scale, diversity, annotation approaches, and domain coverage. The study identified persistent challenges including class imbalance, cultural bias, and insufficient contextual variation, concluding that broader data collection across languages, demographics, and real-world scenarios is essential for developing more robust emotion-aware systems.[3]

A Review on EEG-Based Multimodal Learning for Emotion Recognition: A review focusing on EEG-based multimodal emotion recognition explored the integration of physiological brain signals with behavioral indicators such as facial expressions and speech. EEG data was presented as a more objective measure of emotional states given its direct reflection of neural activity. Various fusion architectures combining EEG with other modalities using CNNs and LSTMs were examined, along with preprocessing strategies for signal noise reduction. Findings indicated that multimodal fusion considerably enhances detection accuracy, with meaningful applications in mental health monitoring, adaptive learning, and affective computing. [4]

Deep Emotion Recognition in Textual Conversations: A Survey:T A survey of deep learning methods applied to emotion recognition within textual conversations provided a thorough examination of architectures including CNNs, RNNs, Bi-LSTMs, and transformer-based models capable of capturing contextual dependencies across dialogue turns. Key challenges identified included sarcasm detection, emotional transitions between conversational turns, and contextual ambiguity. Comparative analysis across datasets such as DailyDialog, EmotionLines, and MELD revealed that transformer-based architectures like BERT and RoBERTa consistently achieved superior performance owing to their capacity to model long-range dependencies and subtle emotional shifts throughout conversations.[5]

6. EXPERIMENTAL ANALYSIS

A Twitter emotion dataset encompassing several emotional categories was used for experimental evaluation. Prior to model training, the dataset was cleaned by removing noise such as URLs, user mentions, and special characters, then divided into training and testing portions using an 80:20 split. DistilBERT was applied to generate contextual sentence-level representations from tweet content through its [CLS] token output. These feature embeddings were subsequently fed into a Logistic Regression classifier for emotion categorization, which was selected due to its computational efficiency and strong generalization capability across high-dimensional input spaces.

Model performance was assessed using Accuracy, Precision, Recall, and F1-score as standard evaluation metrics. Experimental findings demonstrated that the hybrid DistilBERT and Logistic Regression system consistently surpassed conventional TF-IDF based machine learning approaches, while also maintaining significantly lower computational demands in comparison to fully fine-tuned transformer models. These results confirm the overall practicality and effectiveness of the proposed hybrid framework for emotion recognition from tweet data.

EVALUATION PARAMETERS

The effectiveness of the proposed tweet emotion detection framework is assessed using four standard classification metrics, namely Accuracy, Precision, Recall, and F1-score. Accuracy reflects the overall correctness of predictions made by the model. Precision quantifies the proportion of correctly identified emotion labels relative to all predicted instances, whereas Recall captures the model's capability to successfully detect actual emotional occurrences within the dataset. The F1-score, computed as the harmonic mean of Precision and Recall, offers a comprehensive and balanced assessment of overall system performance, particularly in scenarios where class distribution is uneven

7. CONCLUSION

recent advancements in hybrid deep learning frameworks, combining handcrafted linguistic features with transformer-based architectures such as BERT, RoBERTa, and XLNet, have significantly improved emotion detection capabilities in social media text. The incorporation of attention mechanisms, transfer learning, and fine-tuning strategies has collectively enhanced model accuracy, generalization, and sensitivity to subtle emotional cues across diverse datasets and domains. These developments represent a substantial leap forward in enabling intelligent systems to accurately interpret and respond to human emotions, with promising applications spanning sentiment monitoring, social media analysis, and human-computer interaction, ultimately paving the way for more personalized and emotionally aware AI-driven solutions.

REFERENCES

[1] P. Kulkarni, Y. Kale, P. Ahire, A. Dayma, and S. Berad, "Detection of Fake Online Reviews Using Machine Learning and Removal of Fake Reviews," Journal of Engineering, Computing & Architecture, vol. 13, no. 4, 2022.

[2] S. R. Basha, M. S. B. Rao, P. K. K. Reddy, and G. R. Kumar, "Emotional Tweets Analysis on Social Media with Short Text Classification Using Various Machine Learning Techniques," Journal of Computational and Theoretical Nanoscience, vol. 17, pp. 1–6, Dec. 2020, doi: 10.1166/jctn.2020.9442.

[3] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, "Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages," Sustainability, vol. 15, no. 16, p. 12539, Aug. 2023, doi: 10.3390/su151612539.

[4] S.Al-Azami and E.-S,M. E1-Alfy, “A Review and Critical Analysis of Multimodal Datasets for Emotional AI,” Artificial Intelligence Review,vol. 58, 2025.

[5] R. Pillalamarri and U. Shanmugam, “A Review on EEG-Based Multimodal Learning for Emotion Recognition,” Artificial Intelligence Review , vol. 58, 2025.

[6] P. Pereira, H. Moniz, and J.P. Carvalho,“Deep Emotion Recognition in Textual Conversations: A Survey,” Artificial Intelligence Review, vol. 58, 2025.

[7] T. Chutia and N. Baruah, “A Review on Emotion Detection Using Deep Learning,” Artificial Intelligence Review, vol. 57, 2024.

[8] Z. Liu, K.Yang, Q. Xie, T. Zhang, and S. Ananiadou, “EmoLLMs: Emotional Large Language Models and Annotation Tools for comprehensive affective analysis ,” Proc.ACM SIGKDD Conf. Knowledge Discovery and Data mining (KDD),2024.

[9] S. Wu, X. Wang,L.Wang,D.He, and J. Dang, “Enriching Multimodal Sentiment Analysis via Emotional Descriptions of visual-audio content,” arXiv preprint arXiv:2412.10460, 2024.

[10] T. Parvin and M. M. Hoque, “An Ensemble Technique to Classify Multi-Class Textual Emotion,” Procedia Computer Science , vol.193, pp. 72-81,2021.

[11] V. KP, R. AB, G. HL,V.Ravi and M. Krichen, “A Tweet Sentiment Classification Approach Using an Ensemble Classifier,” International Journal of Coginitive Computing in Engineering , vol. 5, pp. 170-177,2024.