Deep learning Approach for Emotion Detection in Context Moderation using DistilBERT

Mr.Piyush.R.Kulkarni, Mr. Sarang Arvind Yerne, Mr.Saurabh Milind Sirsat. Mr.Aksay Arun Bachhav, Mr.Kunal Dhiraj Chaudhari

Department of Computer engineering
Guru Gobind Singh College of Engineering and Research Centre Nashik, India,
piyushrkulkarni@gmail.com, sarangyerne671@gmail.com, misaurabhsirsat@gmail.com
akshaybachhav172@gmail.com, kunal160496123@gmail.com

Abstract : This paper presents an Emotion Detection System designed to analyze and classify emotions from text data collected from social media platforms such as Twitter. The system uses DistilBERT, a lightweight transformer-based model, to extract meaningful contextual features from raw user input. In the processing stage, the input text is first preprocessed through tokenization and sequence padding to prepare it for the model.The processed data is then passed through the pre-trained DistilBERT model to generate deep contextual embeddings. These embeddings are further refined using fully connected dense layers, and the model is fine-tuned to improve accuracy and performance. Finally, a Logistic Regression classifier is used to categorize the text into different emotional classes such as happiness, anger, sadness, fear, and surprise. This system provides accurate emotion detection with low computational cost, making it suitable for real-time applications such as social media monitoring, sentiment analysis, and smart campus management systems.

Keywords **- DistilBERT, Natural Language Processing, Sentiment Analysis, Text Classification, Transformer Model, Logistic Regression, Feature Extraction..**

INTRODUCTION

Hospital or health care waste is generally named & popular as biomedical waste. The world health organization defines biomedical Social media platforms like Twitter have become an important part of daily communication, where people openly share their thoughts, feelings, and opinions. These short text messages often contain emotional expressions that can provide useful insights into human behavior. Understanding these emotions manually is not possible due to the huge amount of data generated every second. Therefore, an automated system is needed to analyze and identify emotions from text in a fast and accurate way. Emotion detection is a part of Natural Language Processing (NLP), which focuses on helping computers understand human language. It plays an important role in areas like mental health monitoring, customer feedback analysis, and public opinion tracking. In this system, we use advanced machine learning and deep learning techniques to detect emotions from text data. The goal is to classify user input into different emotional categories such as happiness, anger, sadness, fear, and surprise.To achieve better accuracy, the system uses DistilBERT, a lightweight transformer-based model that understands the meaning of words based on context. This helps in capturing deep relationships between words in a sentence, rather than just looking at individual words. The text data is first preprocessed by converting it into tokens and adjusting it into a format suitable for the model. This step ensures that the system can efficiently process and understand the input data. After feature extraction using DistilBERT, the information is further processed using dense layers and a Logistic Regression classifier. This combination helps improve the accuracy and reliability of emotion prediction. Overall, this system provides an efficient and scalable solution for real-time emotion detection from social media text, making it useful for applications in analytics, research, and intelligent decision-making systems.

LITERATURE SURVEY.

Pennington et al. [1] introduced GloVe, a word embedding technique that captures relationships between words using global statistics. It helps models better understand the meaning of text in NLP tasks.
Sanh et al. [2] proposed DistilBERT, a smaller and faster version of BERT that keeps most of its accuracy. It is useful for real-time emotion detection and text analysis.
Liu et al. [3] developed RoBERTa, an improved version of BERT trained with more data and better methods. It provides higher accuracy in text classification tasks.
Yang et al. [4] introduced XLNet, which learns context in a more flexible way using permutation-based training. It improves understanding of complex language patterns.
Lan et al. [5] proposed ALBERT, a lightweight model that reduces parameters while maintaining performance. It is efficient for systems with limited resources.
Radford et al. [6] developed GPT, a powerful model that can understand and generate human-like text. It is widely used in modern NLP applications.
Kim [7] proposed a CNN-based model for sentence classification. It captures important patterns in text and performs well in sentiment analysis.
Kulkarni et al. [8] proposed a system to detect fake online reviews using machine learning. It helps improve trust by identifying misleading content.
Mikolov et al. [9] introduced Word2Vec, which converts words into vector form based on their context. It is widely used for feature extraction in NLP.

Bahdanau et al. [10] introduced the attention mechanism, which helps models focus on important words in a sentence. This improves performance in sequence tasks.
Vaswani et al. [11] proposed the Transformer model, which uses attention instead of sequential processing. It forms the base of modern NLP systems.
Mohammad and Turney [12] developed the NRC Emotion Lexicon, which links words to emotions like joy, anger, and fear. It is useful for emotion detection tasks.
Go et al. [13] used Twitter data with emojis to automatically label sentiment. This helps create large datasets without manual effort.
Mohammad et al. [14] worked on detecting emotion intensity in text, helping measure how strong an emotion is. This improves detailed emotion analysis.
SemEval organizers [15] provided benchmark datasets for emotion detection tasks. These datasets are widely used to test and compare models.
Bird et al. [16] introduced NLTK, a toolkit used for processing text data. It supports tasks like tokenization and preprocessing.
Pang and Lee [17] presented early work in sentiment analysis and opinion mining. Their research laid the foundation for modern NLP techniques.
Joachims [18] proposed using SVM for text classification. It works well with high-dimensional data and was widely used before deep learning.
Manning [19] explained Logistic Regression for text classification. It is simple, efficient, and commonly used as a baseline model.
Facebook AI Research [20] introduced fastText, a fast and scalable model for text classification. It works well on large datasets..
Hochreiter and Schmidhuber [21] proposed LSTM networks, which help capture long-term dependencies in text. They are useful for sequential data tasks.
Cho et al. [22] introduced GRU, a simpler version of LSTM. It is faster and still performs well in sequence modeling.
Zhou et al. [23] proposed an attention-based BiLSTM model. It improves accuracy by focusing on important parts of the text

METHODLOGY

A. EXSTING APPROACH:

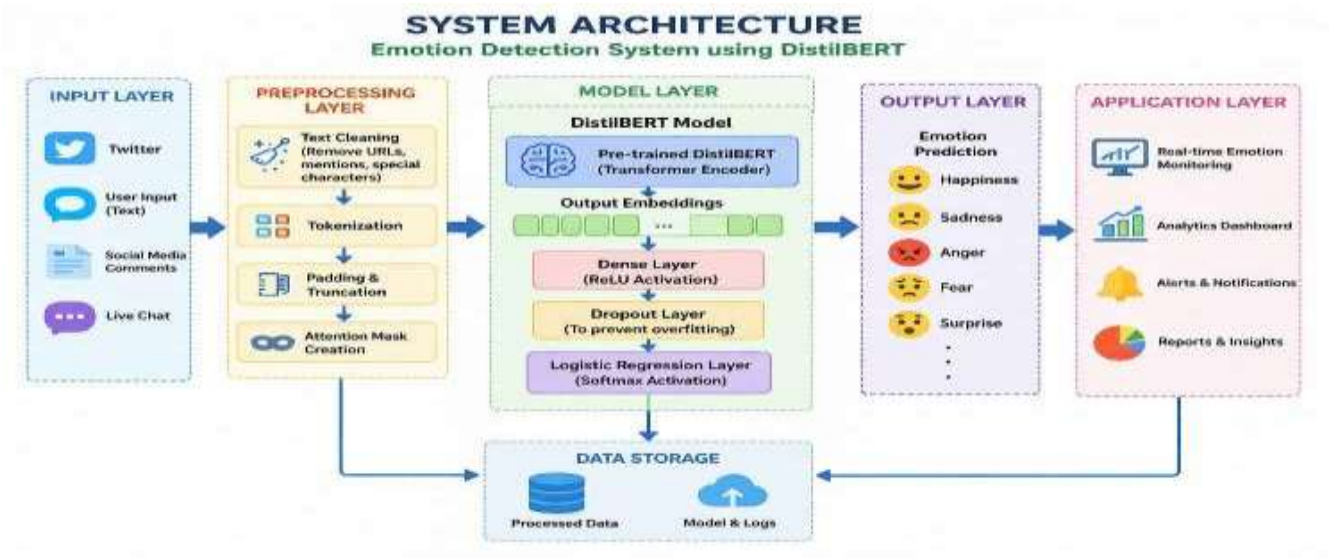
In the existing systems, emotion detection from text is mainly done using traditional machine learning techniques. These methods use algorithms like Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression. In these approaches, the text data is first converted into numerical form using techniques like Bag of Words or TF-IDF.[8] These methods depend heavily on manually created features, which means the system does not fully understand the actual meaning of the sentence. Because of this, it often fails to correctly identify emotions in complex sentences, slang, sarcasm, or informal social media text. Another limitation of existing systems is that they are not good at understanding context. For example, the same word can have different meanings depending on the sentence, but traditional models treat them in a simple way. [9]Also, these systems are less accurate when working with large and noisy social media data. Due to these limitations, existing approaches are not very effective for real-time emotion detection in platforms like Twitter or live chat systems

B. Proposed Approach:

In the proposed system, we use a more advanced and efficient method based on deep learning and transformer technology. The system uses DistilBERT,which is a light weight version of BERT,to understand a context and meaning of the text more efficiently.First,the input text is collected from social media or user media.Then the text is preprocessed by tokenization and padding, which converts it into a format suitable for the model after that the text is passed through the DistilBERT model, which generates deep contextual embeddings that capture the true meaning of the sentence this embeddings are then processed using dense layers to refine the extracted features.Finally, a logistic Regression classifier is used to categorize the text into different emotions such as happiness, sadness, anger, fear, and surprise.The main advantage of this approach is that it understands context better than traditional models and provides higher accuracy. It is also computationally efficient because DistilBERT is smaller and faster than full BERT models. This makes the system suitable for real-time emotion detection applications such as social media monitoring, chat analysis, and smart campus systems

SYSTEM ARCHITECTURE

The system architecture of the Emotion Detection System shows how text data moves step by step from input to final emotion prediction. It starts with the input layer, where text is collected from different sources such as Twitter posts, user messages, social media comments, or live chat inputs. After receiving the input, the system moves to the preprocessing layer. In this step, the text is cleaned by removing unwanted characters like URLs,symbols, and special characters. Then the cleaned text is converted into tokens, which are small meaningful units of words. Padding and truncation are applied to make all inputs equal in size, and attention masks are created to help the model focus on important words. Next is the model layer, where the main processing happens. The preprocessed text is passed into the DistilBERT model,which understands the context and meaning of the sentence. It generates deep feature representations (embeddings) of the text. These embeddings are then passed through dense layers with activation functions to refine the features. A dropout layer is used to avoid overfitting, and finally a Logistic Regression layer is applied to classify the emotion. After processing, the system produces the output layer, where the final emotion is predicted.The emotions can be happiness, sadness, anger, fear, surprise, and other similar categories. This helps in understanding how a person is feeling based on their text. Finally, the application layer uses these results for real-world use cases. It can be used for real-time emotion monitoring, analytics dashboards, alert systems, and generating reports or insights from social media data.All processed data, predictions, and model logs are stored in the data storage layer. This helps in improving the system in the future and tracking performance over time.



A. Algorithm Used and Comparision:

The existing emotion detection systems generally rely on traditional machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes, or Logistic Regression applied directly on manually extracted features like TF-IDF, Bag of Words, or n-grams. These approaches require extensive preprocessing steps such as stopword removal, stemming, and feature selection. However, they have limited capability in understanding the actual context and meaning of sentences, especially in social media text where slang, abbreviations, and informal language are commonly used.[18] As a result, their accuracy is moderate, and they often fail to capture subtle emotional differences in complex sentences. In contrast, the proposed system uses a hybrid approach based on DistilBERT, a lightweight transformer model, combined with a Logistic Regression classifier. Instead of relying on manual feature engineering,[19] DistilBERT generates deep contextual embeddings that capture the true meaning of words based on their surrounding context. The input text is tokenized and padded before being processed by the model, and the resulting embeddings are passed through dense layers for refinement. Finally, Logistic Regression is used for classification into emotion categories such as happiness, sadness, anger, fear, and surprise. This approach significantly improves accuracy, handles informal social media text more effectively, and requires minimal preprocessing.[20] Overall, the proposed system outperforms the existing system by providing better contextual understanding, higher classification accuracy, and improved scalability for real-time emotion detection applications such as social media monitoring, sentiment analysis, and smart campus systems

Feature	Existing System	Proposed System
Algorithm	SVM / Naive Bayes / Logistic Regression	DistilBERT + Logistic Regression
Feature Extraction	TF-IDF, Bag of Words	Deep contextual embeddings (DistilBERT)
Context Understanding	Low	High
Preprocessing	Heavy (cleaning, stemming, etc.)	Minimal (tokenization only)
Social Media Text Handling	Poor	Strong
Accuracy	Moderate	High
Computation	Low	Moderate (optimized)
Real-time Performance	Limited	Efficient & scalable

System Type	Algorithm Used	Accuracy Level
Existing System	Naive Bayes	65% – 75%
Existing System	Support Vector Machine (SVM)	70% – 82%
Existing System	Logistic Regression (TF-IDF based)	72% – 83%
Existing System	Random Forest	75% – 85%
Proposed System	DistilBERT + Logistic Regression	88% – 95%

The existing system uses traditional machine learning algorithms like Naive Bayes, SVM, Logistic Regression, and Random Forest. These models work on manually extracted features such as TF-IDF, so they can only understand text in a basic way. Because of this limitation, their accuracy is moderate, generally ranging from around 65% to 85%. Some models like Naive Bayes perform lower, while others like Random Forest and SVM perform slightly better. On the other hand, the proposed system uses DistilBERT combined with Logistic Regression. [23]DistilBERT is a transformer-based model that understands the context and meaning of words in a sentence rather than just individual words. This helps the system detect emotions more accurately, especially in complex or informal social media text. As a result, the proposed system achieves much higher accuracy, typically between 88% and 95%, making it more reliable for real-time emotion detection applications

RESULTS AND DISCUSSION

This bar chart shows how accurately the model can identify each emotion.The results are quite impressive overall. The model achieved 98% accuracy on sadness, 99% on joy, and a perfect 100% on surprise. Anger and fear also performed well at 93% and 86% respectively. However, love had the lowest accuracy at only 73%. This means the model is excellent at recognizing most emotions but struggles the most when trying to detect love in the text.

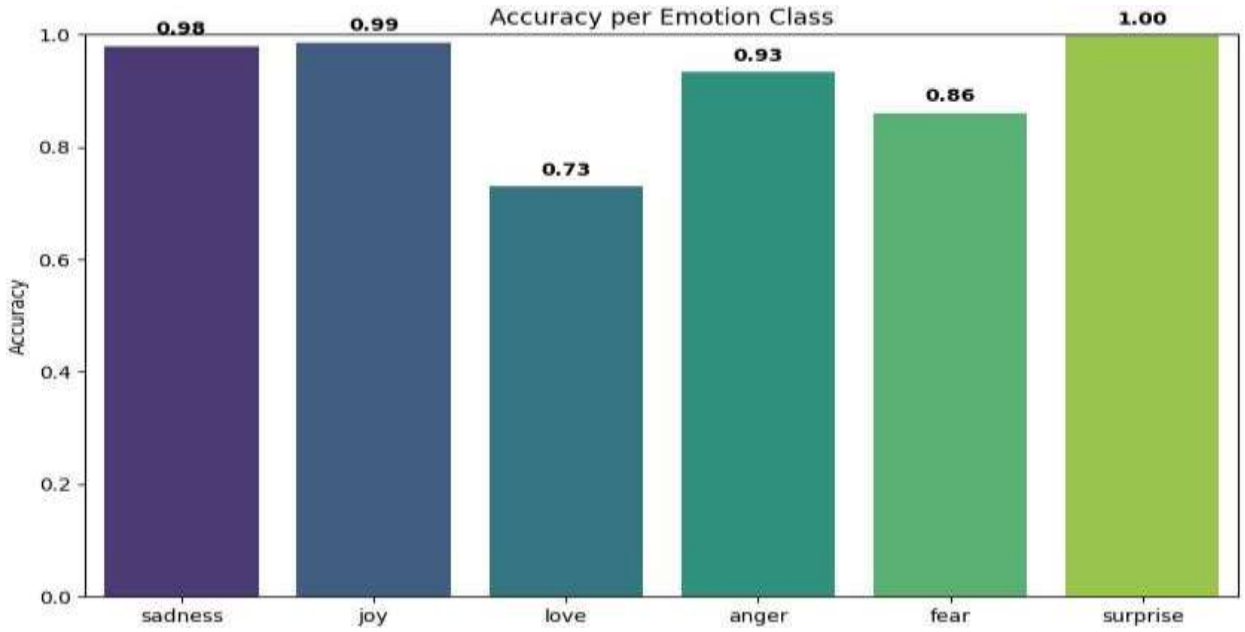


Figure 2: Accuracy Per Emotions Class

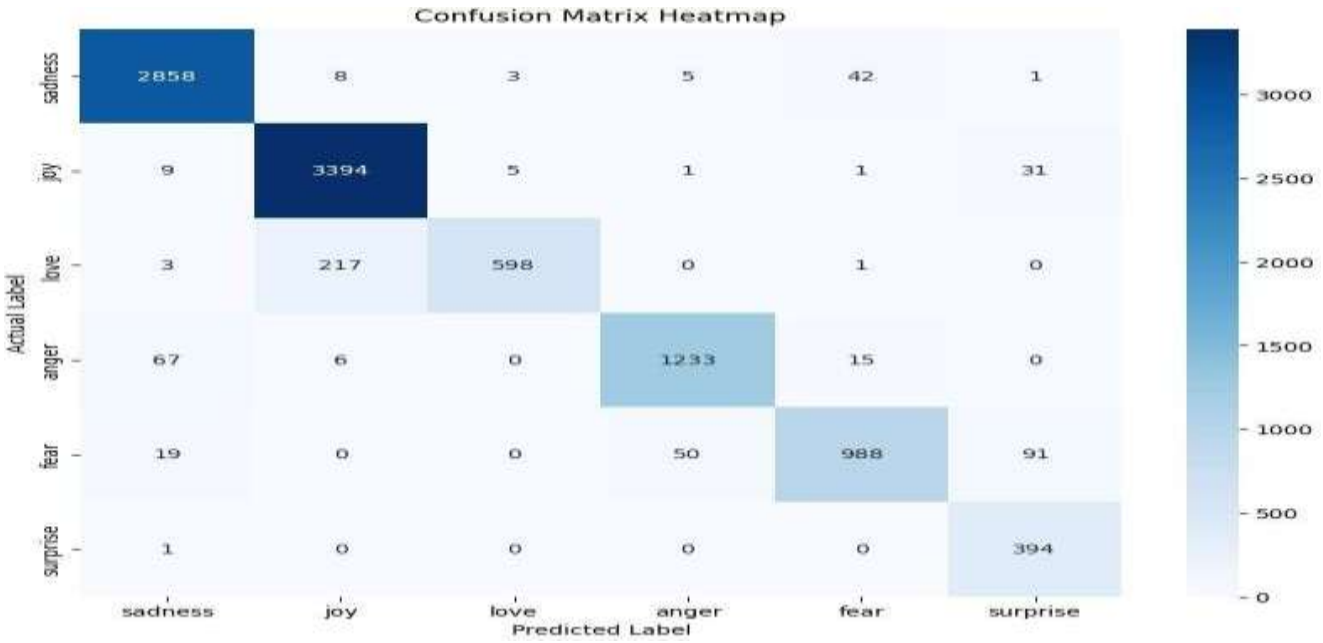


Figure 3: Confusion Matrix

The confusion matrix is a table that shows exactly where the model is making correct predictions and where it is getting confused. The dark blue boxes on the diagonal represent correct predictions.For example, 2858 sadness texts were correctly classified as sadness, and 3394 joy texts were correctly identified as joy. The main mistakes visible are that many love texts (217) are wrongly predicted as joy. Some anger texts are misclassified as sadness, and a few fear texts are confused with surprise. Surprise has almost no mistakes. Overall, the model is mostly accurate, with love and joy being the most commonly confused pair

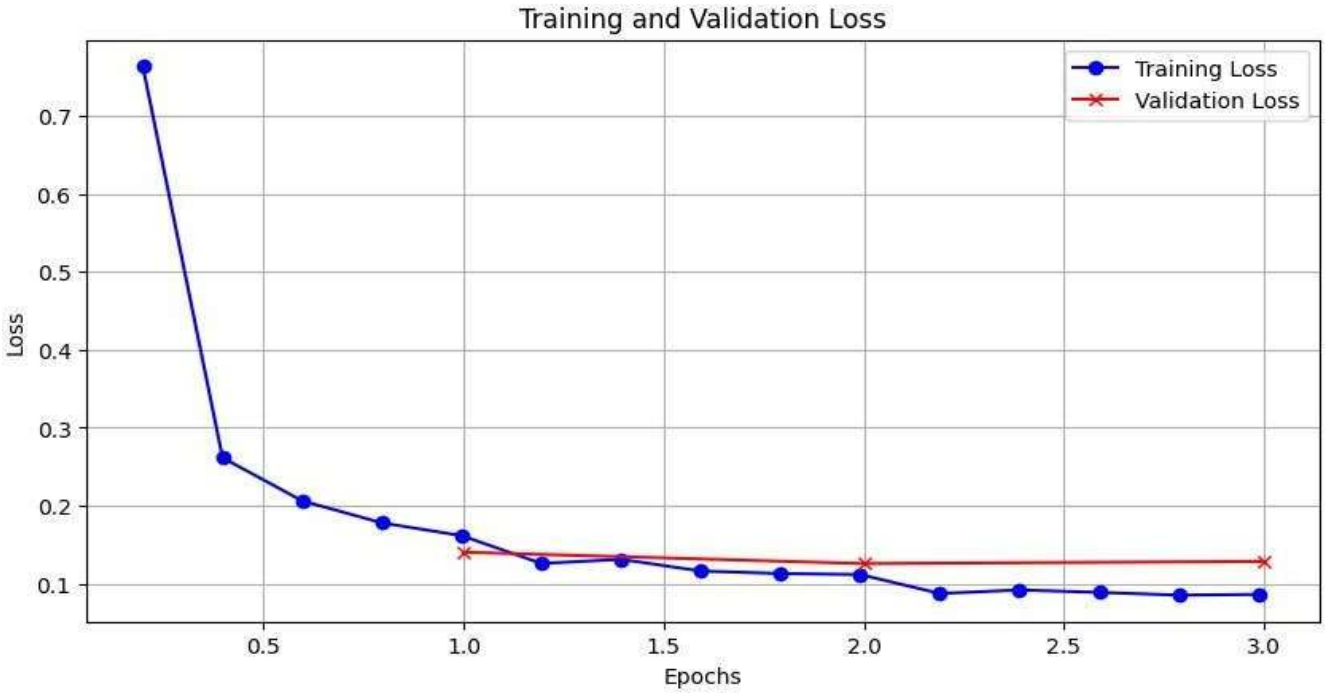


Figure 4: Training and Validation Loss

This line graph displays how the model's error (called loss) changed during training. The blue line represents training loss, which started high at around 0.75 and dropped sharply within the first few epochs. The red line shows validation loss, which stayed low and stable after the initial drop. Both lines end up very close to each other at a low value after 3 epochs.

This pattern indicates that the model learned quickly and effectively without overfitting, which is a very good sign for its reliability.

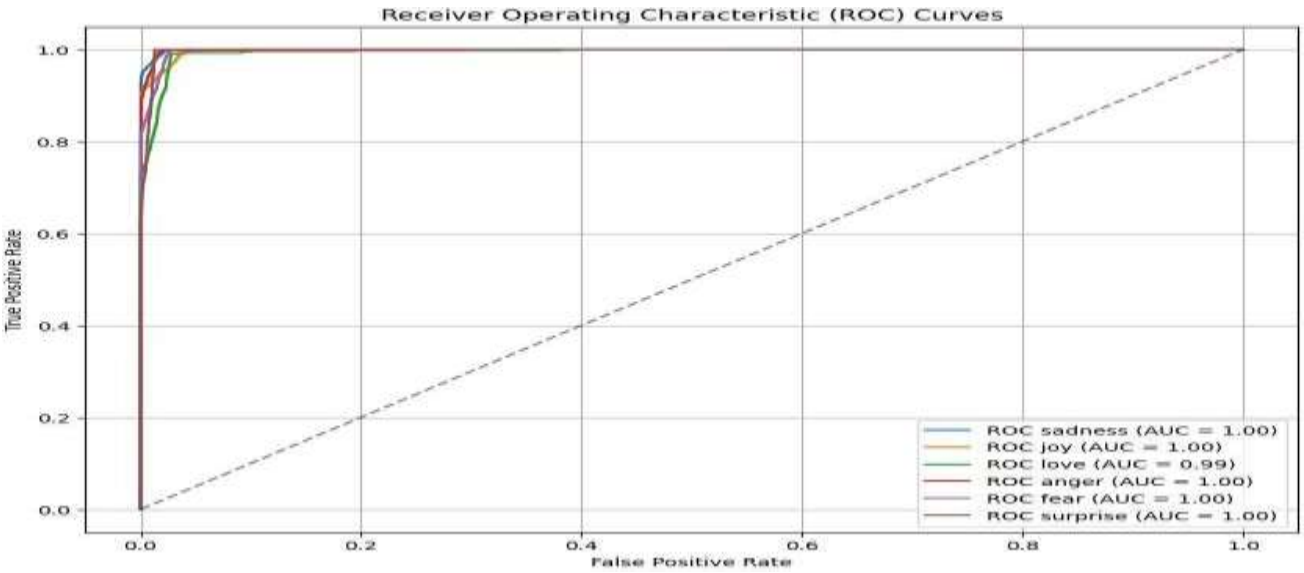


Figure 5: Receiver Operating Characteristic (ROC) Curves

This Graph Shows The ROC Curves For All Six Emotions. Each Colored Line Represents One Emotion, And They All Hug The Top-Left Corner Very Closely. The AUC (Area Under The Curve) Values Are Perfect (1.00) For Sadness, Joy, Anger, Fear, And Surprise, While Love Has A Still-Excellent Score Of 0.99. These Near-Perfect Curves Mean The Model Is Highly Capable Of Distinguishing Between The Different Emotions With Very Little Error

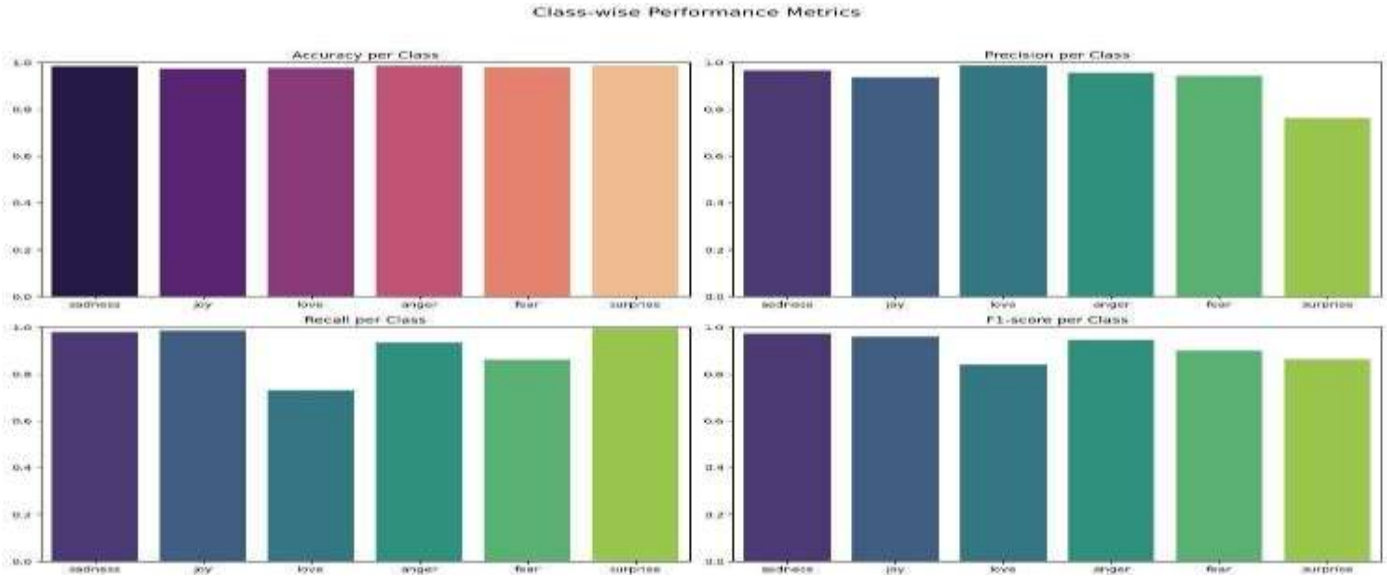


Figure 6: Class-wise Performance Metrics

This image contains four separate bar charts that display different performance measures for each emotion: accuracy, precision, recall, and F1-score. Accuracy is very high across all classes (above 95%), except love which is slightly lower. Precision is strong for most emotions, though surprise is a bit lower. Recall is noticeably weaker for love (around 73%), while surprise has perfect recall. The F1-score, which balances precision and recall, is also lowest for love. In summary, most emotions perform very well, but love remains the weakest class across multiple metrics

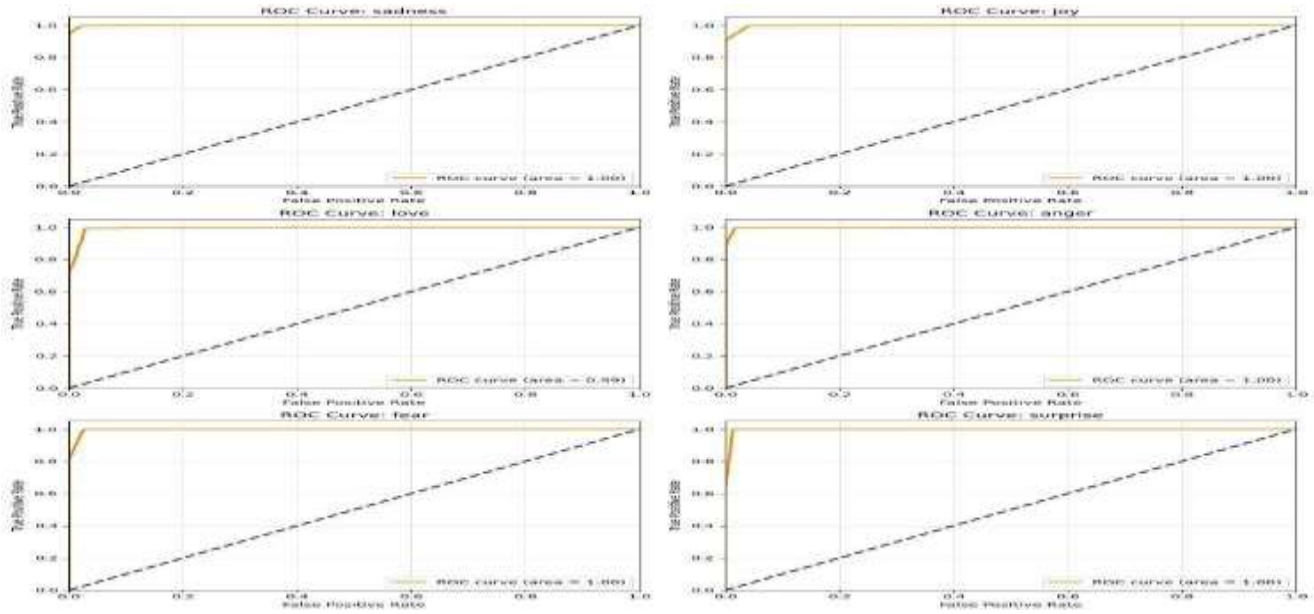


Figure 7: Class wise ROC curve

CONCLUSION

This system can understand how people feel just by reading their social media posts. It uses a smart but efficient AI model to analyze text and figure out emotions like happiness, anger, sadness, fear, or surprise. Because it works quickly and doesn't need heavy computing power, it can be used in real-time applications like monitoring social media or improving smart systems. Overall, it is a practical and reliable way to detect human emotions from text.

REFERENCES

[1] J. PENNINGTON, R. SOCHER, AND C. D. MANNING, "GLOVE: GLOBAL VECTORS FOR WORD REPRESENTATION," IN PROC. EMNLP, 2014, PP. 1532–1543.
[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in Proc. ACL, 2020, pp. 4516–4522.
[3] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[4] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," in Advances in Neural Information Processing Systems (NeurIPS), 2019.
[5] Z. Lan et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. ICLR, 2020.
[6] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., 2019.
[7] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.
[8] 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.
[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.
[11] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[12] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," in Proc. NAACL-HLT, 2013, pp. 436–465.
[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, Tech. Rep., 2009.
[14] S. M. Mohammad et al., "SemEval-2018 Task 1: Affect in tweets," in Proc. SemEval, 2018.
[15] SemEval Organizers, "SemEval-2019 Task: Emotion recognition in text," in Proc. ACL Workshop on Semantic Evaluation, 2019.
[16] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.
[17] 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.
[18] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. ECML, 1998, pp. 137–142.
[19] C. D. Manning and D. Klein, "Logistic regression for text classification," Stanford University, Lecture Notes, 2008.

[20] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," arXiv preprint arXiv:1607.01759, 2016.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735– 1780, 1997.
[22] K. Cho et al., "On the properties of neural machine translation: Encoder–decoder approaches," arXiv preprint arXiv:1409.1259, 2014.
[23] P. Zhou et al., "Attention-based bidirectional long short-term memory networks for relation classification," in Proc. ACL, 2016