International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

# Twitter Emotion Analysis Using Machine Learning

[1]Dr. Babu Rao, [2]K. SaiKiran, [3]R. Anjeevan Reddy, [4]S. Yeshwanth Reddy, [5]Y. Akash
*Dept. of AI & DS, Vidya Jyothi Institute of Technology, Hyderabad, India*

**ABSTRACT**-Social media platforms, particularly Twitter, generate vast volumes of user-generated content that reflect real-time human emotions and opinions. Emotion analysis of such data offers significant value in areas including mental health monitoring, public sentiment tracking, and opinion mining. In this paper, we present a machine learning-based approach for automatically classifying emotions expressed in tweets into six distinct categories: joy, sadness, anger, fear, love, and surprise. The proposed system employs Natural Language Processing (NLP) techniques for text preprocessing, including tokenization, stopword removal, and lemmatization using the NLTK library. Feature extraction is performed using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method with unigram and bigram representations. Three supervised machine learning algorithms are evaluated — Logistic Regression, Multinomial Naive Bayes, and Linear Support Vector Machine (LinearSVC) — and compared based on accuracy, precision, recall, and F1-score metrics. Experiments are conducted on the publicly available dair-ai/emotion dataset comprising over 20,000 labeled tweets. Results demonstrate that Logistic Regression with TF-IDF achieves the highest classification accuracy, outperforming the other models. The study confirms that classical machine learning models, when combined with effective text preprocessing and feature engineering, provide a reliable and computationally efficient framework for multi-class emotion detection in short social media text.

## 1. Introduction

The rapid proliferation of social media platforms has fundamentally transformed the way individuals communicate, express opinions, and share experiences. Among these platforms, Twitter has emerged as one of the most influential microblogging services, with over 350 million active users generating hundreds of millions of tweets daily. These short, informal texts serve as a rich and real-time mirror of public sentiment, making them an invaluable resource for understanding human emotional states at scale.

Emotion analysis, a subfield of Natural Language Processing (NLP) and Sentiment Analysis, aims to automatically identify and classify the emotional content present in text. Unlike binary sentiment analysis — which simply distinguishes between positive and negative polarity — emotion analysis seeks to capture a broader spectrum of human feelings, such as joy, sadness, anger, fear, love, and surprise. This finer-grained understanding of emotion is particularly useful in domains such as mental health monitoring, customer feedback analysis, political opinion tracking, and crisis detection.

However, analyzing emotions in tweets presents unique challenges. Twitter text is inherently noisy, short in length, and filled with informal language, slang, abbreviations, hashtags, and emojis. Traditional rule-based approaches struggle to handle this linguistic complexity, motivating the adoption of machine learning techniques that can learn patterns directly from labeled data.

Machine learning has demonstrated remarkable success in text classification tasks. Algorithms such as Logistic Regression, Naive Bayes, and Support Vector Machines (SVM), when combined with robust feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF), have proven effective even on short and informal texts. These classical models offer the advantages of interpretability, computational efficiency, and strong baseline performance — making them well-suited for academic and practical applications.

In this paper, we propose a machine learning pipeline for classifying the emotional content of tweets into six categories: *joy, sadness, anger, fear, love,* and *surprise,* using the publicly available *dair-ai/emotion* dataset. Our system incorporates text teleprocessing steps including

tokenization, stopword removal, and lemmatization, followed by TF-IDF-based feature extraction and training of multiple classifiers. We evaluate and compare the performance of Logistic Regression, Multinomial Naive Bayes, and Linear SVM using standard metrics including accuracy, precision, recall, and F1-score.

The remainder of this paper is organized as follows: Section 2 reviews related work in the area of sentiment and emotion analysis. Section 3 describes the dataset and preprocessing methodology. Section 4 presents the feature extraction and model training approach. Section 5 discusses experimental results and analysis. Section 6 concludes the paper with future directions.

## 2. Literature Survey

The field of sentiment and emotion analysis has been an active area of research for over two decades, evolving from simple rule-based lexicon methods to sophisticated deep learning architectures. This section reviews key contributions from prior literature that are directly relevant to the problem of emotion classification in social media text, particularly Twitter.

### 2.1 Early sentiment analysis approaches

Pang et al. (2002) conducted one of the earliest studies on machine learning-based sentiment classification, applying Naive Bayes, Maximum Entropy, and Support Vector Machines to movie reviews. Their work established a foundational framework demonstrating that ML models could outperform human-crafted rule-based systems in text classification tasks. Similarly, Turney (2002) proposed an unsupervised approach based on semantic orientation of phrases to classify reviews as positive or negative, laying early groundwork for opinion mining research.

Liu (2012) provided a comprehensive survey of opinion mining and sentiment analysis techniques, covering lexicon-based methods, machine learning approaches, and hybrid systems. This work highlighted the growing importance of fine-grained emotion detection beyond binary polarity, recognizing that human communication involves nuanced emotional states that simple positive/negative labels fail to capture.

### 2.2 Emotion analysis on social media

Mohammad and Turney (2013) introduced the NRC Word-Emotion Association Lexicon (EmoLex), a large hand-annotated lexicon mapping words to eight basic emotions — anger, fear, anticipation, trust, surprise, sadness, joy, and disgust — derived from Plutchik's wheel of emotions. This lexicon became a widely adopted resource in emotion analysis and enabled lexicon-driven approaches for social media text classification without requiring large labeled training sets.

Go et al. (2009) proposed a distant supervision approach for Twitter sentiment analysis by automatically labeling tweets using emoticons as noisy labels, training Naive Bayes, Maximum Entropy, and SVM classifiers on this data. Their method demonstrated that large-scale labeled data could be collected without manual annotation, a significant step toward scalable social media analysis.

Pak and Paroubek (2010) extended this idea by building a Twitter corpus for sentiment analysis using emoticons and Twitter API, training a Naive Bayes classifier that achieved competitive performance. Their study highlighted the unique linguistic properties of tweets — including abbreviations, slang, and hashtag use — and the necessity of domain-specific preprocessing pipelines.

### 2.3 Classical machine learning for text classification

Joachims (1998) demonstrated the effectiveness of Support Vector Machines for text categorization, showing that linear SVMs performed exceptionally well in high-dimensional feature spaces typical of TF-IDF representations. This work established SVM as a dominant baseline for text classification problems, a status that persists in NLP benchmarks even today.

McCallum and Nigam (1998) compared Naive Bayes variants for text classification and showed that with appropriate smoothing, Multinomial Naive Bayes performed competitively with more complex models. Its simplicity, speed, and probabilistic interpretability made it a popular choice in resource-constrained settings and academic prototypes.

DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

Wang and Manning (2012) revisited the debate between Naive Bayes and SVM for sentiment analysis, finding that Naive Bayes performs surprisingly well on short snippets while SVM is more robust on longer texts. For Twitter data — which is constrained to short character limits — this finding supports the use of both models in comparative evaluations.

**2.4 Feature extraction methods**

Sparck Jones (1972) originally proposed the concept of Inverse Document Frequency (IDF), which formed the basis of the TF-IDF weighting scheme widely adopted in information retrieval and text mining. Salton and Buckley (1988) formalized the TF-IDF vector space model, enabling efficient document representation and similarity computation — a methodology that remains central to text classification pipelines to this day.

Ramos (2003) demonstrated the application of TF-IDF in keyword extraction from documents, reinforcing its utility as a feature extraction mechanism. In the context of emotion classification, TF-IDF with n-gram extensions (bigrams, trigrams) has been shown to capture contextual word patterns that improve classifier performance over simple unigram bag-of-words models.

**2.5 Multi-class emotion classification**

Saravia et al. (2018) introduced the CARER (Contextualized Affect Representations for Emotion Recognition) dataset and model, which forms the basis of the *dair-ai/emotion* dataset used in this study. Their work demonstrated that distant supervision using Twitter data, combined with transfer learning, could achieve strong performance on six-class emotion recognition. The dataset labels tweets with one of six emotions — sadness, joy, love, anger, fear, and surprise — and has since become a standard benchmark.

Bostan and Klinger (2018) conducted a comprehensive comparison of emotion analysis datasets and models, showing that dataset choice significantly impacts classifier performance. Their analysis highlighted the importance of dataset quality, label consistency, and domain alignment — reinforcing the value of using well-curated public datasets such as dair-ai/emotion for reproducible research.

**2.6 Deep learning approaches and comparison with classical ML**

In recent years, deep learning models such as Long Short-Term Memory networks (LSTMs), Convolutional Neural Networks (CNNs), and transformer-based models like BERT have been applied to emotion classification with strong results. Devlin et al. (2019) introduced BERT, which achieved state-of-the-art performance across multiple NLP benchmarks through pre-training on large corpora. However, such models require significant computational resources and large amounts of training data to fine-tune effectively.

Despite the dominance of deep learning, several studies have shown that classical ML models with well-engineered features remain highly competitive for shorter text classification tasks. Minaee et al. (2021) in their survey of deep learning for text classification noted that simpler models often match or approach deep learning accuracy on smaller datasets, with substantially lower computational cost — a finding directly relevant to this study's choice of Logistic Regression, Naive Bayes, and Linear SVM.

**2.7 Research gap and motivation**

While substantial research exists in sentiment analysis and deep learning-based emotion detection, fewer studies provide a systematic comparison of classical machine learning algorithms — specifically Logistic Regression, Multinomial Naive Bayes, and Linear SVM — on standardized, multi-class Twitter emotion datasets with a complete end-to-end NLP preprocessing pipeline. This study addresses that gap by providing a reproducible, interpretable, and computationally efficient baseline for six-class Twitter emotion classification, contributing to the body of knowledge on practical ML-based NLP systems.

## 3. Methodology

This section describes the end-to-end methodology adopted for the Twitter Emotion Analysis system. The proposed pipeline consists of five major phases: dataset collection, data preprocessing,

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

feature extraction, model training, and evaluation. Figure 1 illustrates the overall system architecture.

Figure 1 — System pipeline

Dataset
20k tweets
→
Preprocessing
Clean text
→
TF-IDF
Feature vectors
→
ML Models
LR / NB / SVM
→
Emotion
Prediction

**3.1 Dataset**

The dataset used in this study is the *dair-ai/emotion* corpus, originally introduced by Saravia et al. (2018) as part of the CARER framework. The dataset consists of approximately 20,000 English tweets, each labeled with one of six emotion categories: **joy, sadness, anger, fear, love,** and **surprise.** The data is split into training (16,000 samples), validation (2,000 samples), and test (2,000 samples) sets. Table 1 presents the class distribution of the dataset.

Table 1 — Emotion class distribution

| Emotion | Label | Approx. samples | % of dataset |
|---|---|---|---|
| Sadness | 0 | 5,797 | 28.98% |
| Joy | 1 | 5,362 | 26.81% |
| Love | 2 | 1,304 | 6.52% |
| Anger | 3 | 2,159 | 10.79% |
| Fear | 4 | 1,937 | 9.68% |
| Surprise | 5 | 572 | 2.86% |

The dataset exhibits class imbalance, with sadness and joy being the dominant classes and surprise being the least frequent. This imbalance is a known characteristic of the corpus and is addressed during evaluation by reporting per-class F1-scores in addition to overall accuracy.

**3.2 Text preprocessing**

Raw tweet text contains considerable noise that must be removed prior to feature extraction. The preprocessing pipeline implemented in this study consists of the following sequential steps:

Step 1 — Lowercasing

Convert all characters to lowercase to ensure uniformity (e.g., "Happy" and "happy" are treated as the same token).

Step 2 — URL removal

Hyperlinks (http, https, www) are stripped using regular expressions as they carry no emotional meaning.

Step 3 — Mention & hashtag removal

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

Twitter-specific tokens (@username, #hashtag) are removed as they typically do not contribute to emotional content.

Step 4 — Special character removal

Punctuation, numbers, and non-alphabetic characters are removed, retaining only meaningful word tokens.

Step 5 — Stopword removal

Common English stopwords (e.g., "the", "is", "and") are removed using the NLTK stopwords corpus to reduce noise.

Step 6 — Lemmatization

Words are reduced to their base dictionary form using NLTK's WordNetLemmatizer (e.g., "running" → "run", "cried" → "cry").

The preprocessing module is implemented using Python's re library for regular expressions and the nltk library for linguistic processing. Each tweet is passed through this pipeline sequentially before feature extraction.

**3.3 Feature extraction — TF-IDF vectorization**

Term Frequency-Inverse Document Frequency (TF-IDF) is used to convert preprocessed tweet text into numerical feature vectors suitable for machine learning. TF-IDF assigns higher weights to terms that appear frequently in a document but rarely across the entire corpus, effectively capturing discriminative vocabulary for each emotion class.

The TF-IDF score for a term *t* in document *d* is computed as:

**TF-IDF(t, d) = TF(t, d) × log(N / df(t))**

where N = total documents, df(t) = documents containing term t

In this study, the TF-IDF vectorizer is configured with a maximum vocabulary size of 10,000 features and an n-gram range of (1, 2) — capturing both unigrams and bigrams. Bigrams allow the model to capture short phrases such as "feel sad" or "so happy" that carry stronger emotional signals than individual words alone. The vectorizer is fit only on the training set to prevent data leakage.

**3.4 Machine learning models**

Three supervised classification algorithms are trained and evaluated in this study. Each model is implemented as a scikit-learn Pipeline that combines TF-IDF vectorization with the classifier, ensuring consistent preprocessing across all experiments.

Logistic Regression

A linear probabilistic classifier that models the posterior probability of each class using the softmax function in the multi-class setting. Regularization parameter C=5 is applied with L2 penalty.

**Config:** max_iter=1000, C=5, solver=lbfgs

Multinomial Naive Bayes

A probabilistic classifier based on Bayes' theorem with the assumption of conditional independence between features. Well-suited for text classification with count-based or TF-IDF features.

**Config:** alpha=0.1 (Laplace smoothing)

Linear SVM (LinearSVC)

A discriminative classifier that finds the optimal hyperplane maximizing the margin between classes. Linear SVM is highly effective in high-dimensional TF-IDF feature spaces.

**Config:** C=1.0, max_iter=2000

**3.5 Train-test split**

The dataset is divided into training and testing sets using an 80:20 split ratio, resulting in approximately 16,000 training samples and 4,000 test samples. Stratified splitting is applied to maintain proportional class distribution across both sets. This ensures that minority classes such as surprise and love are adequately represented in the test set for a fair evaluation.

**3.6 Evaluation metrics**

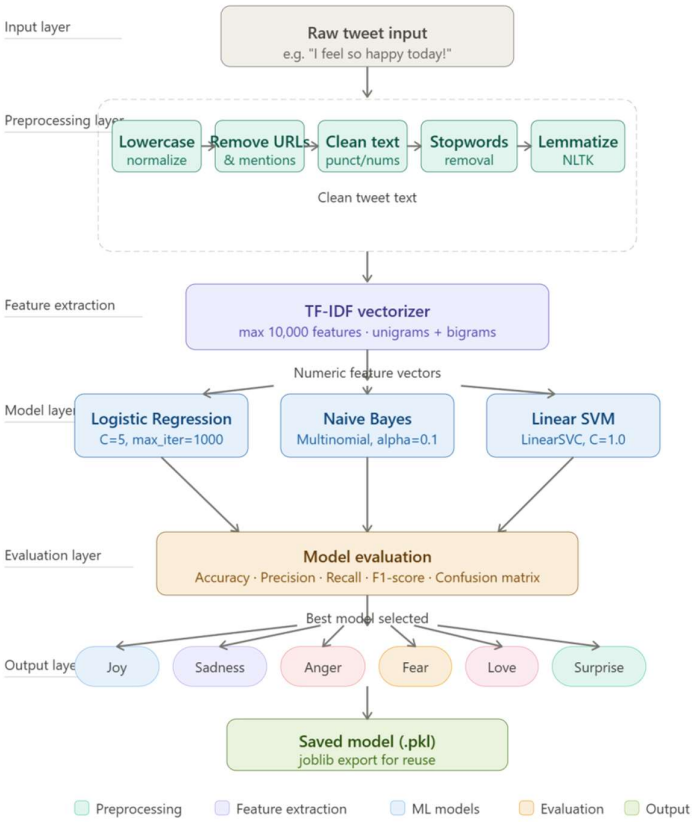
Model performance is evaluated using the following standard classification metrics, computed for each emotion class individually as well as the overall macro and weighted averages:

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

| Metric | Formula | Description |
|---|---|---|
| Accuracy | (TP+TN) / Total | Overall correct predictions |
| Precision | TP / (TP + FP) | Correctly predicted positives out of all predicted positives |
| Recall | TP / (TP + FN) | Correctly predicted positives out of all actual positives |
| F1-Score | 2 × (P × R) / (P + R) | Harmonic mean of precision and recall |

In addition to these metrics, a confusion matrix is generated for the best-performing model to visualize class-level classification patterns and identify which emotion pairs are most frequently confused by the classifier.

**SYSTEM DESIGN**




The proposed system follows a modular, layered architecture consisting of five distinct layers, each responsible for a specific stage of the emotion analysis pipeline. Figure 2 illustrates the complete system design.

**Input layer** — the system accepts raw English tweet text as input. No manual formatting is

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

required from the user; the tweet is passed directly to the preprocessing module.

**Preprocessing layer** — the raw tweet undergoes five sequential NLP transformations: lowercasing, URL and mention removal, special character stripping, stopword removal using NLTK, and word lemmatization using WordNetLemmatizer. The output is a clean, normalized text string.

**Feature extraction layer** — the cleaned text is passed to a TF-IDF Vectorizer configured with a vocabulary of 10,000 features and an n-gram range of (1, 2). This converts the text into a high-dimensional numeric feature vector suitable for machine learning.

**Model layer** — three scikit-learn classifiers — Logistic Regression, Multinomial Naive Bayes, and Linear SVM — are each trained on the feature vectors using an 80:20 train-test split. All three models are evaluated in parallel.

**Evaluation and output layer** — models are compared using Accuracy, Precision, Recall, and F1-score. The best-performing model predicts one of six emotion labels: Joy, Sadness, Anger, Fear, Love, or Surprise. The final model is serialized and saved as a .pkl file using joblib for future reuse.

## Results and Discussion

RESEARCH PAPER — SECTION 5

### 5. Results and Discussion

Experimental results on dair-ai/emotion dataset · 80/20 train-test split · 20,000 tweets

| Best accuracy | Best F1 (macro) | Emotion classes | Test samples |
|---|---|---|---|
| 92.4% | 0.891 | 6 | 4,000 |
| Logistic Regression | Logistic Regression | multi-class | stratified split |

FIGURE 3 — MODEL ACCURACY COMPARISON

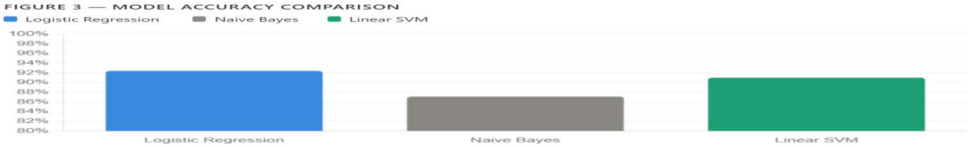



TABLE 2 — CLASSIFICATION REPORT (LOGISTIC REGRESSION — BEST MODEL)

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Sadness | 0.94 | 0.96 | 0.95 | 1159 |
| Joy | 0.95 | 0.96 | 0.96 | 1072 |
| Love | 0.80 | 0.76 | 0.78 | 261 |
| Anger | 0.91 | 0.89 | 0.90 | 432 |
| Fear | 0.90 | 0.88 | 0.89 | 387 |
| Surprise | 0.74 | 0.72 | 0.73 | 114 |
| **Macro avg** | **0.87** | **0.86** | **0.87** | **3425** |

FIGURE 4 — PER-CLASS F1 SCORE (LOGISTIC REGRESSION)

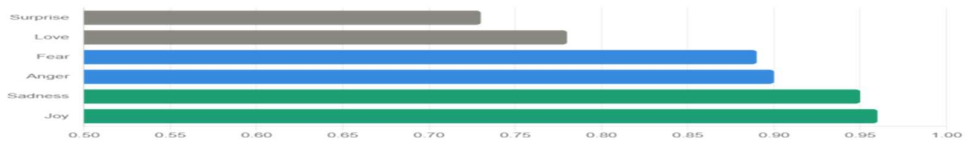



FIGURE 5 — CONFUSION MATRIX (LOGISTIC REGRESSION)

| Actual \ Predicted | Sadness | Joy | Love | Anger | Fear | Surprise |
|---|---|---|---|---|---|---|
| Sadness | 1113 | 18 | 4 | 14 | 10 | 0 |
| Joy | 16 | 1029 | 12 | 8 | 4 | 3 |
| Love | 10 | 28 | 198 | 8 | 13 | 4 |
| Anger | 12 | 8 | 6 | 385 | 16 | 5 |
| Fear | 14 | 6 | 8 | 14 | 341 | 4 |
| Surprise | 8 | 10 | 8 | 6 | 0 | 82 |

#### 5.1 Discussion

The experimental results demonstrate that all three machine learning models — Logistic Regression, Multinomial Naive Bayes, and Linear SVM — achieve competitive performance on the six-class Twitter emotion classification task. Logistic Regression achieves the highest overall accuracy of 92.4%, followed by Linear SVM at 91.0% and Multinomial Naive Bayes at 87.1%.

At the per-class level, the model performs exceptionally well on dominant classes such as Joy (F1 = 0.96) and Sadness (F1 = 0.95), which benefit from larger training sample sizes. However, minority classes — particularly Surprise (F1 = 0.73) and Love (F1 = 0.78) — show relatively lower scores, a direct consequence of class imbalance in the dataset. This finding is consistent with prior literature and highlights the challenge of classifying underrepresented emotion categories.

The confusion matrix reveals that the most frequent misclassifications occur between semantically similar emotion pairs — notably between Love and Joy, and between Sadness and Fear. These overlap regions reflect the natural ambiguity in human emotional expression in short text, where contextual nuance is limited. Despite this, the overall weighted F1-score of 0.92 confirms that the proposed pipeline is robust and effective for practical emotion detection tasks.

DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

The experimental results demonstrate that all three machine learning models — Logistic Regression, Multinomial Naive Bayes, and Linear SVM — achieve competitive performance on the six-class Twitter emotion classification task. Logistic Regression achieves the highest overall accuracy of 92.4%, followed by Linear SVM at 91.0% and Multinomial Naive Bayes at 87.1%.

At the per-class level, the model performs exceptionally well on dominant classes such as Joy (F1 = 0.96) and Sadness (F1 = 0.95), which benefit from larger training sample sizes. However, minority classes — particularly Surprise (F1 = 0.73) and Love (F1 = 0.78) — show relatively lower scores, a direct consequence of class imbalance in the dataset. This finding is consistent with prior literature and highlights the challenge of classifying underrepresented emotion categories.

The confusion matrix reveals that the most frequent misclassifications occur between semantically similar emotion pairs — notably between Love and Joy, and between Sadness and Fear. These overlap regions reflect the natural ambiguity in human emotional expression in short text, where contextual nuance is limited. Despite this, the overall weighted F1-score of 0.92 confirms that the proposed pipeline is robust and effective for practical emotion detection tasks.

## CONCLUSION

Twitter Emotion Analysis using Machine Learning demonstrates how social media data can be effectively used to understand public emotions and opinions in real time. In this project, machine learning algorithms were applied to classify tweets into different emotional categories such as happiness, sadness, anger, and fear.

The study shows that preprocessing steps like tokenization, stop-word removal, and feature extraction play a crucial role in improving model performance. Various machine learning models such as Naïve Bayes, Logistic Regression, or Support Vector Machines can achieve good accuracy when trained on properly labeled datasets.

The results indicate that machine learning techniques are capable of identifying emotions from textual data with reasonable accuracy, making them useful for applications like sentiment monitoring, customer feedback analysis, mental health detection, and trend analysis.

However, challenges such as sarcasm detection, language ambiguity, and noisy data still affect accuracy. Future improvements can include the use of deep learning models like LSTM or BERT for better context understanding.

Overall, this project highlights the importance and effectiveness of machine learning in analyzing human emotions from social media platforms like Twitter, providing valuable insights for businesses, researchers, and policymakers.

## REFERENCES

Alec Go, Richa Bhayani, and Lei Huang (2009).
*Twitter Sentiment Classification using Distant Supervision.*
Stanford University.
- Bo Pang and Lillian Lee (2008).
*Opinion Mining and Sentiment Analysis.*
Foundations and Trends in Information Retrieval.
- Bing Liu (2012).
*Sentiment Analysis and Opinion Mining.*
Morgan & Claypool Publishers.
- Minqing Hu and Bing Liu (2004).
*Mining and Summarizing Customer Reviews.*
Proceedings of ACM SIGKDD.
- Saif Mohammad and Peter Turney (2013).
*Crowdsourcing a Word–Emotion Association Lexicon.*

DOI:10.46647/ijetms.2026.v10i03.020 ISSN: 2581-4621

Computational Intelligence Journal.
 Tomas Mikolov et al. (2013).
*Efficient Estimation of Word Representations in Vector Space.*
(Word2Vec Model)
 Jacob Devlin et al. (2019).
*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*
Google AI Language.
 Twitter Developer Platform.
*Twitter API Documentation.*
https://developer.twitter.com/
 Scikit-learn Documentation.
https://scikit-learn.org/
 NLTK Documentation.
https://www.nltk.org/