A Comparative Sentiment Analysis on Tweets Using Simple and Complex Models

Md Enayet Ali Labib
CSE Department
BRAC University
Dhaka, Bangladesh
enayet.ali.labib@g.bracu.ac.bd

Niamul Hasan Chowdhury
CSE Department
BRAC University
Dhaka, Bangladesh
niamul.hasan.chowdhury@g.bracu.ac.bd

Shifat Sharif
CSE Department
BRAC University
Dhaka, Bangladesh
shifat.sharif@g.bracu.ac.bd

***Abstract*—Automatic sentiment and emotion analysis of social media text has become increasingly important for understanding public opinions and emotonal trends. This study performs sentiment and emotion classification on Twitter data using a publicly available Kaggle dataset annotated with sentiment polarity and multiple emotions, including joy, anger, fear, and trust. We evaluate a spectrum of models, from traditional machine learning approaches such as Naive Bayes classifiers to deep learning models including LSTM and Bidirectional RNN, analyzing their ability to handle the noisy and informal nature of tweets. Inspired by benchmark studies such as SemEval-2018 and distant supervision techniques, we compare the performance of simple and complex models, demonstrating that contextual embeddings and affective feature integration substantially improve classification accuracy. Our findings highlight best practices for effective sentiment and emotion analysis on social media content and suggest directions for future research in multi-class and context-aware sentiment modeling.**

I. Introduction

Social media platforms have become critical channels for expressing opinions, emotions, and reactions to current events, products, and social issues. Among these platforms, Twitter is particularly influential due to its high volume of short messages, known as tweets. The sheer number of tweets generated daily makes manual sentiment and emotion analysis impractical.

Sentiment analysis involves classifying text as expressing positive, negative, or neutral opinions, while emotion analysis aims to detect specific emotional states such as joy, anger, fear, or trust. Automatic analysis of tweets is challenging because of informal language, abbreviations, slang, emojis, hashtags, and frequent spelling variations.

In this study, we perform sentiment and emotion classification using a publicly available Kaggle dataset containing tweets labeled with sentiment polarity and multiple emotion categories. We implement and compare a range of models, from traditional Bag-of-Words–based classifiers to deep learning architectures utilizing word embeddings and contextual representations. By evaluating these models, we examine the impact of richer text representations on classification performance, particularly in handling the informal and noisy nature of tweet text.

The results demonstrate that advanced models outperform traditional approaches, effectively capturing contextual and emotional nuances in tweets. This work highlights the potential of machine learning to support applications such as public opinion monitoring, brand analysis, and social trend assessment.

II. Literature Review

Sentiment and emotion analysis on social media has received significant attention in recent years, mainly due to the rapid growth of user-generated content on platforms such as Twitter. Early research in this field focused primarily on sentiment classification, where the goal was to determine whether a piece of text expresses a positive, negative, or neutral opinion. These early approaches typically relied on polarity-based methods and Bag-of-Words features. While these approaches provided useful baseline performance, they were limited in capturing contextual relationships, subtle emotional variations, and the informal language patterns commonly found in tweets.

To address these challenges, more recent studies have shifted toward fine-grained affect modeling and deep learning methods. One of the most influential contributions in this direction is the SemEval-2018 Task on “Affect in Tweets” [4], which introduced benchmark datasets annotated for multiple affective dimensions, including emotion intensity, sentiment valence, and multi-label emotion categories. The shared task demonstrated that traditional polarity analysis is insufficient for understanding emotional nuances, especially in short and noisy social-media text. Many of the top-performing systems in this challenge adopted neural network architectures, word embeddings, and hybrid feature combinations, showing that richer text representations can significantly improve classification performance.

Another important line of work explores the use of distant supervision to automatically construct large-scale sentiment and emotion datasets [2]. Instead of relying solely on manual annotation, some studies use weak signals such as emojis or hashtags to infer emotional labels. These weakly labeled datasets allow researchers to train deep learning models on millions of tweets at a much lower cost. Findings from such studies indicate that recurrent neural networks and pretrained embeddings perform particularly well when trained on large and diverse datasets, and they can generalize effectively to real-world emotion classification tasks.

In addition to deep learning approaches, several works have investigated the role of feature engineering in emotion detection [1]. Comparative studies evaluating lexical features, TF-IDF representations, contextual embeddings, and transformer-based models suggest that hybrid feature strategies often provide strong results. These studies emphasize that although neural architectures are highly effective at learning contextual and semantic representations, handcrafted linguistic features remain useful for capturing affective cues such as sentiment-bearing terms, word frequency patterns, and topic-related indicators. In particular, feature-engineered baselines have shown competitive performance in scenarios involving limited training data or highly imbalanced class distributions. Furthermore, carefully designed lexical and structural features can enhance model robustness when operating on noisy, informal, or short text segments, such as tweets, where contextual signals may be sparse or fragmented. As a result, many recent approaches combine traditional feature engineering with deep representations to leverage complementary strengths from both design paradigms.

Studies examining social-media communication during major social events, such as public crises, transportation environments, and health-related contexts, demonstrate that emotional expressions often evolve over time and correlate with psychological and situational factors. In particular, a recent study on multimodal mental-health prediction using behavioral and digital signals [5] proposes a deep-learning framework that integrates speech patterns, keystroke dynamics, physiological indicators, and behavioral logs to detect early signs of stress, anxiety, and depression. The study shows that fusing multiple models provide richer emotional insight than single-source analysis, while also highlighting ethical and privacy challenges associated with emotion-aware systems. In addition to these methodological advances, several studies have examined the application of sentiment and emotion analysis in real-world contexts, such as public opinion monitoring, social behavior analysis, and crisis communication. For example, research on COVID-19 related Twitter communication [6] has shown that emotional expression in social media posts changes over time and is influenced by major global events, social stress, and public uncertainty. Such findings highlight the importance of emotion-aware text analysis for understanding collective social responses.

More recently, sentiment and emotion analysis has also been applied to geopolitical and information-warfare contexts. A large-scale study of state-sponsored influence operations on Twitter analyzes over two million tweets attributed to campaigns linked to China, Iran, and Russia (Paper Mentioned Here). The authors identify distinct emotional and linguistic strategies used by different state actors: Russian operations primarily employ negative sentiment and toxic language to intensify polarization, Iranian campaigns combine antagonistic and supportive tones to shape ideological alignment, while Chinese activity favors positive or neutral rhetoric to promote strategic narratives. This work demonstrates how emotion, sentiment polarity, and abusive language function as deliberate communication tools in coordinated propaganda ecosystems, extending affective computing research beyond user-driven expression toward strategic narrative engineering.

Overall, the existing literature suggests three major trends in sentiment and emotion analysis on Twitter. First, fine-grained emotion modeling provides deeper insight than simple polarity classification. Second, neural network-based and embedding-based models generally outperform traditional machine learning approaches. Third, combining contextual representations with complementary features enhances robustness when working with noisy social-media text.

Based on these findings, the present study builds upon prior work by evaluating both traditional classifiers and deep learning models for sentiment and emotion classification on tweets, with the aim of analyzing how different model families respond to informal text structure, contextual variation, and affective content.

III. Methodology

A. Dataset

The dataset used in this study is a publicly available Twitter sentiment and emotion corpus from Kaggle, titled “Sentiment and Emotions of Tweets”. The dataset contains English tweets annotated with both sentiment polarity (positive, negative, neutral) and associated emotion categories. Tweets in the dataset reflect real-world social media language characteristics, including informal expressions, abbreviations, hashtags, emojis, and spelling variations.

A total of 24,970 tweets are included in the sentiment classification subset. The class distribution is imbalanced, with negative sentiment being the most frequent category. The distribution of sentiment labels is presented in Table I.

TABLE I
Sentiment Distribution in the Dataset

Sentiment	Number of Tweets
Negative	10,556
Positive	7,366
Neutral	7,048

In addition to sentiment polarity, the dataset also contains emotion categories associated with tweets (e.g., joy, anger, fear, sadness, trust, surprise). To better understand the interaction between sentiment polarity and emotional expressions, a bar chart was generated that visualizes the frequency of different emotions across the three sentiment classes (Negative, Positive, Neutral). This visualization highlights patterns where negative tweets are more frequently associated with emotions such as anger and sadness, while positive tweets are dominated by emotions such as joy and trust.

This dual-label structure allows the dataset to support both sentiment classification and emotion-aware text analysis, making it suitable for evaluating models across varying levels of affective granularity.

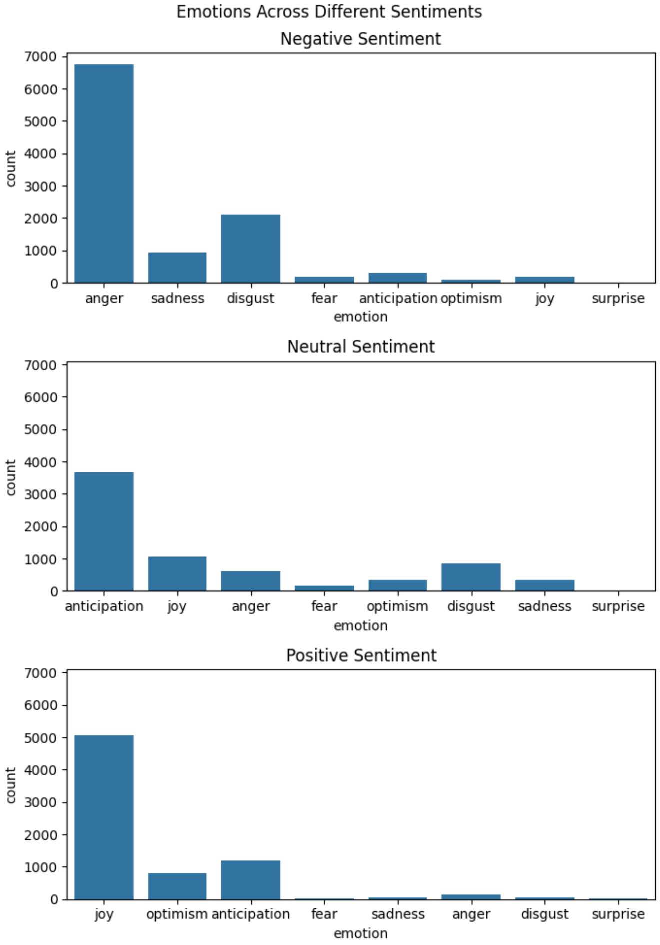

Fig. 1. Emotion Across Different Sentiments

B. Data Preprocessing

Since tweets contain noisy, informal, and user-generated text, preprocessing was applied to normalize and clean the data prior to model training. The following preprocessing steps were performed:

1) Tokenization and optional stopword removal to standardize textual units.
2) Regular expression (RegEx) filtering to remove:
   - punctuation marks,
   - special characters,
   - Twitter handles (`@username`),
   - hashtag symbols (#) while retaining the hashtag keywords.
3) Conversion of all text to lower-case.
4) Removal of extra whitespace and non-alphanumeric characters.

These steps reduce noise while preserving meaningful linguistic and emotional content, particularly in hashtags, which often carry contextual or affective significance in social media text.

C. Models

To evaluate the effect of model complexity on sentiment classification performance, a range of traditional and deep learning models were implemented. The models include classical machine learning classifiers as well as recurrent neural network architectures.

1) Classical Machine Learning Models: The following feature-based models were trained:

- Gaussian Naive Bayes classifier,
- K-Nearest Neighbors (KNN),
- Random Forest classifier.

Tweets were converted into numerical feature representations using Term Frequency–Inverse Document Frequency (TF–IDF) and Bag-of-Words encoding. These models serve as strong performance baselines for comparison with neural network models.

2) Recurrent Neural Network Models: To examine the capability of sequential deep learning architectures in capturing contextual dependencies in tweets, the following models were implemented:

- Simple Recurrent Neural Network (RNN),
- Bidirectional RNN (Bi-RNN).

The Bi-RNN processes textual features in both forward and backward temporal directions, enhancing the ability to model contextual relationships within short tweet sequences.

3) LSTM-Based Deep Learning Model: A Long Short-Term Memory (LSTM) network was trained to handle long-term dependencies and to mitigate vanishing-gradient limitations associated with standard RNNs. Tweets were encoded into padded sequences prior to training, and dropout layers were applied to reduce overfitting.

The inclusion of RNN, Bi-RNN, and LSTM models enables a structured comparison across progressively deeper sequential architectures, providing insights into how increasing model complexity influences sentiment classification accuracy on noisy Twitter text.

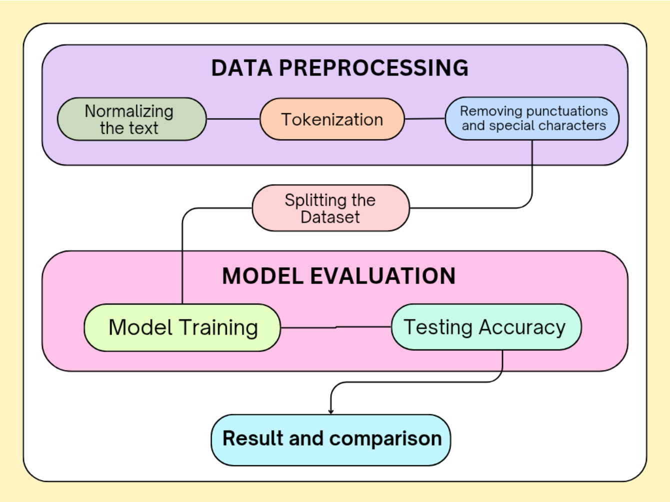

Fig. 2. Methodology Flowchart

IV. Results and Interpretation

The performance of all implemented models was evaluated using accuracy as the primary metric. Table 1 summarizes the obtained results for sentiment and emotion classification on the Twitter dataset.

The LSTM model achieved the highest accuracy (0.7501), demonstrating its effectiveness in modeling sequential dependencies and contextual relationships within short and noisy tweet text. Its ability to retain long-range contextual cues allows it to better capture emotional tone and polarity, which is particularly important for tweets containing slang, abbreviations, or implicit expressions of sentiment.

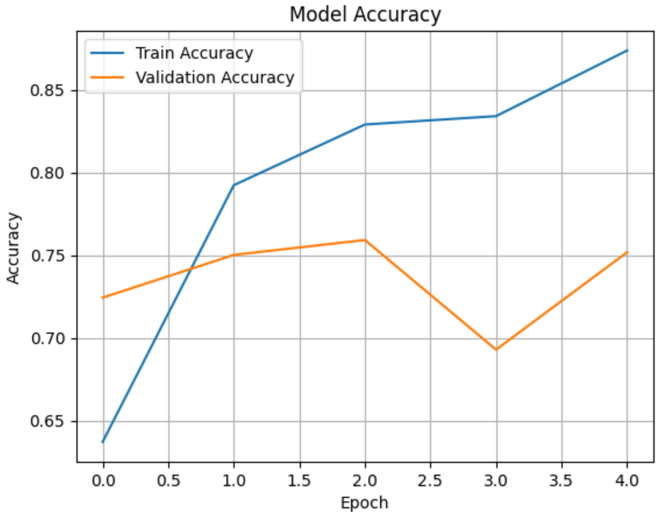

Fig. 3. LSTM Accuracy

The Simple RNN (0.7331) and Bi-Directional RNN (0.7201) also performed competitively, outperforming all traditional classifiers. Although the bidirectional variant enables processing of contextual signals in both forward and backward directions, its accuracy remains slightly lower than the Simple RNN and LSTM. This suggests that while bidirectional context contributes to improved representation, the absence of memory gates limits its ability to handle longer-range contextual patterns compared to LSTM.

Among the traditional models, **Random Forest achieved the best performance (0.7263)**, benefiting from its ensemble-based learning structure and robustness to feature noise. However, its reliance on sparse Bag-of-Words features restricts its capacity to capture semantic and affective nuances in informal text, which explains its lower performance relative to neural models.

In contrast, **Gaussian Naive Bayes (0.6552)** and **K-Nearest Neighbors (0.6600)** produced the lowest accuracies. These results highlight the limitations of statistical independence assumptions in Naive Bayes and the sensitivity of KNN to high-dimensional sparse spaces. Both models struggle to generalize over highly variable and informal linguistic expressions commonly present in tweets.

A separate bar chart was generated to analyze the distribution of emotion categories across sentiment classes. The visualization shows that negative tweets are predominantly associated with emotions such as anger and fear, while positive tweets express joy and trust more frequently. Neutral tweets exhibit a relatively balanced emotion distribution with a lower affective intensity. This observation supports the hypothesis that emotional features reinforce the polarity of feelings and can enhance classifier performance when effectively represented.

A comparative bar chart was generated to visually illustrate the variation in accuracy across models. The results indicate a clear performance hierarchy between traditional machine learning approaches and neural network–based architectures.

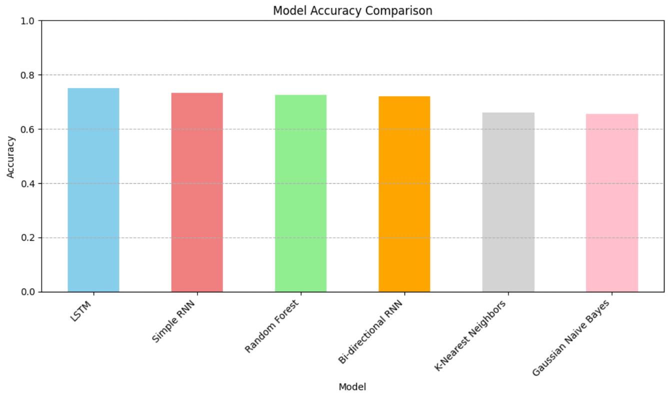

Fig. 4. Model F1 Comparison

Overall, the experimental findings demonstrate that **deep learning models, particularly LSTM, are better suited for the analysis of emotion and social media sentiment**. Their ability to model sequential patterns and contextual semantics enables improved handling of noisy, short-length, and affect-rich text, leading to superior classification performance compared to traditional machine learning approaches.

V. Conclusion

This study investigated sentiment and emotion classification on Twitter data using a publicly available Kaggle dataset and a range of machine-learning and deep learning models. The objective was to analyze how different models perform on noisy, informal tweet text and to examine whether richer textual representations improve classification accuracy. The dataset contained tweets labeled with sentiment polarity and emotion categories, and preprocessing steps such as tokenization, removal of special characters, and handling of hashtags were applied to standardize the text before model training.

The experimental results demonstrate that deep learning models performed more effectively than traditional machine-learning approaches for sentiment classification. Among the evaluated models, the LSTM model achieved the highest accuracy (0.7501), followed by the Simple RNN (0.7331) and Random Forest (0.7263). The Bidirectional RNN also performed almost similarly but was slightly lower than the simple RNN. The Gaussian Naive Bayes and K-Nearest Neighbors obtained the lowest accuracy scores. These findings indicate that models capable of capturing sequential and contextual dependencies in text are better suited for analyzing short and informal social-media messages compared to purely feature-based classifiers.

The results further reinforce observations from prior literature that embedding-based and neural architectures are more effective than Bag-of-Words style approaches for affective computing tasks on Twitter. The superior performance of LSTM and RNN models suggests that the ability to retain temporal and contextual information plays an important role in distinguishing sentiment polarity and emotional cues within tweets. On the other hand, the relatively weaker performance of Naive Bayes and KNN highlights the limitations of shallow models when dealing with informal language, slang, abbreviations, and emotionally expressive content.

Overall, this study demonstrates the usefulness of machine learning and deep learning methods for automated sentiment and emotion analysis on Twitter data. The findings suggest that deep neural models provide more robust and reliable performance for real-world sentiment classification tasks, particularly when working with noisy social-media text. The work contributes to the growing body of research on social-media analytics by providing a comparative evaluation of multiple model families on a unified dataset and by illustrating the importance of contextual representation learning in sentiment and emotion classification.

References

[1] Black, J. T., & Shakir, M. Z. (2025). Emotion on the edge: An evaluation of feature representations and machine learning models. Natural Language Processing Journal, 10, 100127. https://doi.org/10.1016/j.nlp.2025.100127

[2] Kastrati, M., Kastrati, Z., Shariq Imran, A., & Biba, M. (2024). Leveraging distant supervision and deep learning for Twitter sentiment and emotion classification. Journal of Intelligent Information Systems, 62, 1045–1070. https://doi.org/10.1007/s10844-024-00845-0

[3] Khemani, B., Patil, S., Malave, S., & Gupta, J. (2025). Improved graph convolutional network for emotion analysis in social media text. MethodsX, 14, 103325. https://doi.org/10.1016/j.mex.2025.103325

[4] Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018) (pp. 1–17). Association for Computational Linguistics. https://aclanthology.org/S18-1001/

[5] Sayed, M. A., Hossain, M. A., Rahman, M. M., Ali, G. G. M. N., Islam, M. A., Paul, K. C., & Qin, X. (2025). Public sentiment analysis of roadway work zones using social media data and machine learning models. Data Science and Management. https://doi.org/10.1016/j.dsm.2025.04.001

[6] Storey, V. C., & O'Leary, D. E. (2022). Text analysis of evolving emotions and sentiments in COVID-19 Twitter communication. Cognitive Computation, 16, 1834–1857. https://doi.org/10.1007/s12559-022-10025-3