International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05752
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Ensemble Learning with Transformer Models for Sentiment Analysis on Cryptocurrency Tweets

Dr. Y. Mohammed Iqbal[1], S. Mohamed Fawaz[2], Dr. S. Peerbasha[3], Dr. M. Mohamed Surputheen[4], Dr. M. Rajakumar[5]

Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

----------***----------

Abstract- Cryptocurrency markets are strongly influenced by public sentiment shared on social media platforms, especially Twitter, where opinions and reactions can rapidly affect market behaviour. Accurate sentiment analysis of cryptocurrency-related tweets is therefore important for market analysis and decision-support systems. This study presents a comparative analysis of three transformer-based language models—RoBERTa, DeBERTa, and FinBERT—for multi-class sentiment classification of cryptocurrency tweets. To ensure reliable evaluation, experiments are conducted using three train–test split configurations: 60–40, 70–30, and 80–20. The dataset is systematically cleaned, normalized, labelled, and balanced using two preprocessing techniques to support consistent experimentation. Model performance is evaluated using accuracy, precision, recall, F1-score, specificity, and ROC-AUC metrics. In addition to individual model evaluation, a weighted soft-voting ensemble framework is proposed to combine the probabilistic outputs of all models. Experimental results demonstrate that the ensemble approach consistently outperforms the individual models across all evaluation settings, achieving an F1-score of 85.82% and an accuracy of 86.00%, with the best results obtained using the 80–20 split. These findings indicate that ensemble learning improves prediction stability, reliability, and generalization for cryptocurrency sentiment analysis.

Keywords: Cryptocurrency Sentiment Analysis, Transformer Models, RoBERTa, DeBERTa, FinBERT, Ensemble Learning.

1. INTRODUCTION

The rapid expansion of cryptocurrency markets has reshaped modern financial systems, increasing the use of digital platforms for information sharing and investment decision-making. Twitter is one of the platforms that investors, analysts, and the public frequently use to express opinions and feedback regarding cryptocurrency assets. According to previous research [1], Twitter sentiment can have an impact on market behaviour and assist in predicting short-term financial movements. In financial analytics and decision-support systems, automated sentiment analysis of cryptocurrency-related tweets has therefore gained increasing importance. However, due to their informality and context-dependent nature, cryptocurrency-related tweets are challenging to analyse. Social media content often uses abbreviations, hashtags, emojis, colloquialisms, and evolving domain-specific terms. Handcrafted features and basic classifiers in traditional machine learning methods often fail to capture these patterns, limiting their generalization capabilities [2]. As a result of these difficulties, the use of Deep Learning techniques able to better model meaningful and relevant relationships have emerged. Recent developments in natural language processing (NLP), especially transformer-based models, have improved sentiment classification in various text domains. Transformer architectures, which use self-attention mechanisms to understand long-range dependencies and contextual meaning, can understand language better than traditional methods [3]. In sentiment analysis tasks, models such as RoBERTa, which is an optimized version of BERT [4], and DeBERTa, which improves

attention and representation learning, have achieved strong results. Pretrained on financial text corpora, FinBERT has also demonstrated strong performance in financial sentiment analysis [5]. Despite the strong representation capabilities of Transformer models, using a single model can lead to unstable evaluation results. The size of the dataset, the partitioning of the data, and the characteristics of the domain all impact the performance of the model. By combining the results of multiple models, previous studies have shown that ensemble learning techniques can improve the stability and reliability of predictions [6]. Ensemble methods that combine multiple deep learning models have shown promising results in cryptocurrency sentiment analysis [7]. Based on these findings, this study compares RoBERTa, DeBERTa, and FinBERT for multi-class sentiment classification of cryptocurrency-related Twitter data. The outputs of the individual models are combined using a weighted soft-voting ensemble framework. The outputs of the individual models are combined using a weighted soft-voting ensemble framework. To ensure the accuracy of the evaluation, a various of train-test split configurations are used. This work aims to improve classification accuracy, stability, and generalization in crypto-sentiment analysis by combining Ensemble Learning with Transformer-based models.

2. PROBLEM STATEMENT

In sentiment analysis tasks, Transformer-based language models have shown strong performance. However, there are several obstacles to applying these to Twitter data related to cryptocurrency. Tweets about cryptocurrency tend to be informal and context-dependent, often using abbreviations, emojis, sarcasm, and terminology specific to the sector. These features Increasing noise and ambiguity in these features further increases the difficulty of sentiment classification. For evaluation, many existing studies use a single Transformer architecture and a fixed train-test split. Because model performance is sensitive to data partitioning, dataset size, and class distribution, this approach may lead to unstable or biased results. Since the distribution of data varies over time, the reported results may not apply well to the real world. Furthermore, different Transformer models can capture different linguistic and semantic patterns, but these properties often do not combine when the models are evaluated separately. Another difficulty is accurately comparing multiple transformer architectures. It is difficult to evaluate the relative performance of different models without consistent preprocessing procedures, controlled hyperparameter settings, and evaluation across multiple train-test splits. Although various machine learning applications have shown that ensemble learning reduces prediction variance and improves reliability, its structured integration in cryptocurrency sentiment analysis with transformer models is still underexplored.

Therefore, the main goal of this study is to build a reliable sentiment classification framework for cryptocurrency-related Twitter data that: (i) examines a variety of transformer architectures in different train-test split configurations, (ii) ensures that performance evaluations are consistent and reproducible, and (iii) Uses a weighted grouping strategy to increase generalization and classification accuracy.

3. LITERATURE REVIEW

Machine learning and deep learning have been used extensively to study sentiment analysis, especially in the financial text and social media domains. Initially, traditional models such as logistic regression, Naive Bayes, and support vector machines (SVM) were used, along with hand-crafted lexical and statistical features. Although these techniques were understandable and computationally efficient, their ability to capture semantic relationships and long-range dependencies in informal text was limited, making them less useful for difficult sentiment tasks [8]. Language modelling was enhanced by the introduction of distributed word representations. Word2Vec and other similar approaches learn dense vector representations of semantic similarity between words [9]. Deep learning models have used these representations to improve their performance in sentiment analysis and financial prediction. Topic modelling-based sentiment analysis was used to predict the stock market by Nguyen and Shirai [10],

emphasizing the importance of social media-based contextual sentiment signals. By using a self-attention mechanism to model contextual dependencies, the Transformer architecture further improved sentiment classification. BERT was introduced by Devlin et al. [11], and it performed well on a variety of NLP tasks. RoBERTa reported high emotion classification accuracy and improved the pretraining process [12]. DeBERTa improved the performance of context-dependent representation and classification by enhancing the attention mechanism [13] Domain-specific Transformer models have demonstrated efficiency in financial applications. FinBERT, developed by ProsusAI [14], was successful in analyzing financial news and market text after pretraining on financial corpora. Khedr et al. [15] emphasized the importance of sentiment indicators in forecasting when discussing the application of machine learning and deep learning models for predicting cryptocurrency prices. Ensemble learning methods have also been used to improve sentiment classification. By combining their predictions, ensemble classifiers can outperform individual models, as reported by Zhang et al. [16]. The LSTM-GRU ensemble framework of Aslam et al. [7] was found to perform better than a single model in cryptocurrency sentiment analysis. Recent advances in deep learning have shown significant success in medical image analysis, particularly for detecting COVID-19 from lung imaging data. In one study, a framework known as COVID Net-Predictor was proposed to improve diagnostic accuracy using chest imaging datasets. The model integrates a multi-head Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) layers to capture both spatial and sequential characteristics of medical images. A hybrid optimization approach was applied to improve feature selection and enhance classification performance. In addition, preprocessing, segmentation, and feature fusion stages were incorporated to improve robustness across diverse datasets. The experimental results demonstrated high predictive accuracy, indicating that the system could support real-time clinical decision making during pandemic situations [19].

Another study presented an optimized deep learning framework for automated COVID-19 prediction using lung imaging modalities such as chest X-rays and CT scans. The proposed approach employed an enhanced CNN architecture combined with transfer learning techniques to improve model generalization, particularly when training data is limited. Data augmentation and noise reduction techniques were applied to improve robustness and reduce overfitting. The framework was able to accurately differentiate COVID-19 cases from normal and other pneumonia conditions. The findings highlight the effectiveness of intelligent imaging-based systems in enabling rapid screening and reducing the diagnostic burden on healthcare professionals [20]. Despite these advancements, many studies use a fixed train-test split to evaluate performance and employ a single model architecture. Particularly in dynamic areas like cryptocurrency markets, these evaluation methods may result in unstable outcomes. Furthermore, the use of multiple transformer models and data split configurations simultaneously has received less attention. To improve classification accuracy and generalization in cryptocurrency sentiment analysis, this study investigates RoBERTa, DeBERTa, and FinBERT under different train-test splits and proposes a weighted soft-voting ensemble framework.

4. DATASET DESCRIPTION

A. Dataset Overview

The experiment used a balanced and curated dataset of tweets about cryptocurrencies collected from Twitter. This dataset is designed for multi-class sentiment classification and contains public opinions and discussions related to the cryptocurrency market. Each tweet is labelled positive, neutral, or negative. The dataset was balanced across the three sentiment classes to guarantee accurate evaluation. The final dataset contains 9,000 tweets, with an equal number of samples for each class. This balanced distribution helps avoid classification bias and makes it easier to compare different models and training-test split settings in a consistent manner. On Kaggle, the dataset

ISSN: 2583-6129
DOI: 10.55041/ISJEM05752

is available to the public: https://www.kaggle.com/datasets/leoth9/crypto-tweets/data.

B. Dataset Properties and Feature Description

In addition to the additional features created during preprocessing, the dataset also includes unstructured text that was collected from social media posts. Each tweet was cleaned using text normalization steps such as lowercasing and noise removal in addition to the original content. To help with the analysis and model, several additional features were extracted. These include a binary indicator indicating the presence of question marks, the total number of tokens in the cleaned text, and the number of exclamation points. Each tweet is labelled positive (2), neutral (1), or negative (0). This dataset combines text content with these additional features to perform in-depth sentiment analysis of cryptocurrency-related discussions. The main characteristics of the dataset are given in Table 1, including label distribution, data type, total number of samples, sentiment categories, and domain focus. The dataset is designed for balanced multi-class sentiment classification and is limited to the cryptocurrency industry.

Table 1: Dataset Properties Table

Attribute	Description
Data Type	Text (Twitter Tweets)
Total Samples	9,000
Sentiment Classes	3 (Positive, Neutral, Negative)
Domain	Cryptocurrency
Label Distribution	Balanced
Additional Features	Token length, punctuation indicators

C. Class Distribution and Feature Correlation Analysis

To further analyse the dataset, a correlation analysis was used to examine the relationship between the sentiment labels and the additional textual features extracted. The resulting heatmap shows in Fig. 1, shows the relationship between sentiment categories and features such as the number of tokens used, punctuation, and the presence of question marks. This analysis justifies the use of these features in the preprocessing stage and helps determine whether they are useful for sentiment classification.

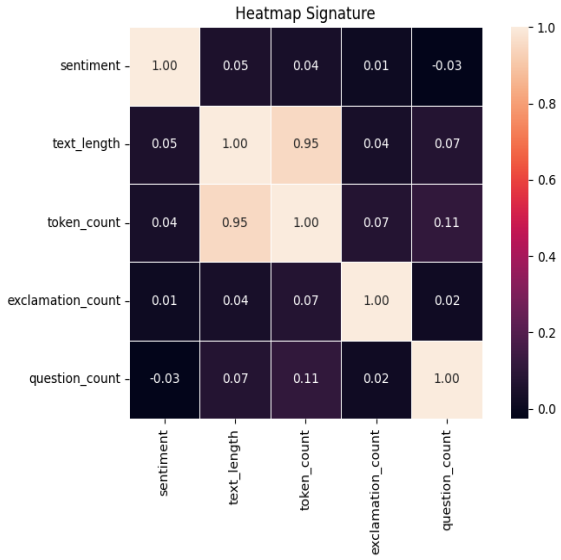

Fig. 1. Correlation heatmap illustrating relationships between sentiment labels and auxiliary textual features in the cryptocurrency tweet dataset.

5. DATA PREPROCESSING AND SENTIMENT LABELING METHODOLOGY

To prepare cryptocurrency-related Twitter data for Transformer-based classification, this study uses a structured and repeatable process for sentiment labelling and preprocessing. Preprocessing is essential to preserve the meaning associated with sentiment as well as improve data quality, as social media text is often informal and noisy. To check the consistency in performance and to see how text normalization and additional feature integration affect the classification results, two different preprocessing methods are used.

A. Dataset Loading and Encoding Detection

The Twitter dataset was created using CSV (comma-separated values). Character encoding may vary or be unknown because social media data often includes special characters, emojis, and text written in multiple languages. To fix this, an automatic encoding detection step using byte-level sampling was implemented. This ensures that the data is loaded correctly and that character corruption won't impact tokenization or introduce semantic errors during preprocessing. Let B be a representation of the raw byte stream of the dataset. The most likely character encoding is identified with the encoding detection function $f_{\text{enc}}(B)$,and the dataset is decoded using this encoding before any further processing. Because incorrect decoding can alter the text content and subsequently negatively impact model performance, this approach is considered preferable to assuming a fixed encoding such as UTF-8.

B. Text Cleaning Strategy

Cryptocurrency tweets often include shortened URLs, user mentions, hashtags, emojis, extra whitespace, and duplicate content. During model training, these elements can introduce irrelevant patterns that do not necessarily reflect sentiment. To manage this, two text normalization approaches were applied: a basic preprocessing pipeline and an advanced pipeline that includes additional feature extraction.

C. Methodology 1: Simple Text Preprocessing

Basic normalization is used in the first preprocessing pipeline to get rid of surface-level noise and retain context information needed by transformer models. Because excessive text modification can alter the relationships that depend on context and are essential for the self-attention mechanism, only minimal cleaning steps are performed.

A tweet should be represented as a series of tokens:

$$T = \{w_1, w_2, \dots, w_n\}$$

Let U, M, and H denote sets of tokens representing URLs, user mentions, and hashtags.

Noise removal is defined as follows:

$$T_1 = \{w_i \in T \mid w_i \notin (U \cup M \cup H)\}$$

In this step, non-informative tokens are removed, but words related to the sentiment remain. Token-level filtering supports transformer-based representations by preserving sentence structure and context instead of removing entire phrases.

Lowercase conversion is then applied:

$$T_2 = \{\text{lower}(w_i) \mid w_i \in T_1\}$$

Where "lower"(.) is converted from alphabetical character to lowercase. This results in better embedding consistency and reduced vocabulary variation due to case differences. Unlike stemming and lemmatization, lowercasing does not change the form of the word, and the meaning remains.

Whitespace normalization is performed as:

$$T_3 = \text{normalize\_space}(T_2)$$

This removes extra spaces without changing the order in which the tokens appear. Finally, empty and duplicate tweets are removed from the dataset.

$$D = \{T_1, T_2, \dots, T_m\}$$

using:

$$D' = \{T_i \in D \mid\mid T_i \mid > 0\} \setminus \text{Duplicates}(D)$$

Removing duplicate entries improves generalization and reduces bias from repeated content without increasing the size of the dataset.

D. Sentiment Labelling and Class Balancing

After text normalization, sentiment labels are assigned using a lexicon-based approach. Each tweet is assigned a polarity score of p, which is then classified into three sentiment classes:

$$y = \begin{cases} 0, & p < 0 (\text{Negative}) \\ 1, & p = 0 (\text{Neutral}) \\ 2, & p > 0 (\text{Positive}) \end{cases}$$

This approach ensures consistent and reusable labelling, without the need for manual annotation. The polarity score provides a consistent value that can be directly mapped to different sentiment categories. To

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05752
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

correct class imbalance, under-sampling is done based on the size of the smallest class. This ensures balanced evaluation across all sentiment categories, while preventing the creation of artificial samples and maintaining original language patterns. The overall process of Methodology 1 is shown in Fig. 2.

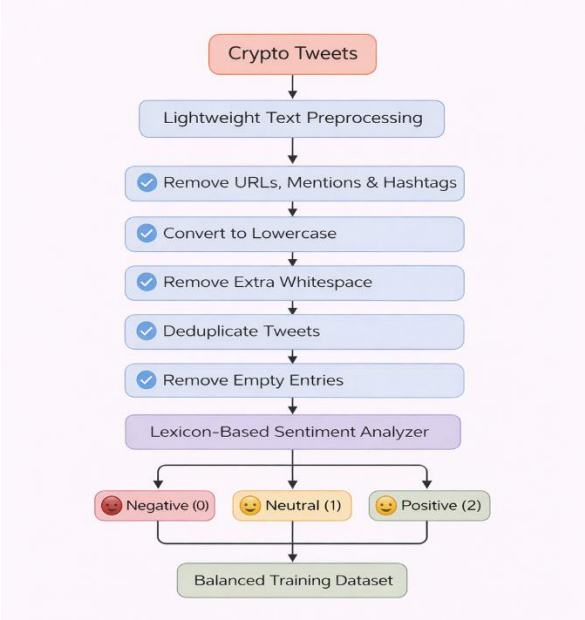

Fig. 2. Workflow of Methodology 1 showing lightweight text preprocessing and lexicon-based sentiment labelling for cryptocurrency tweets.

E. Methodology 2: Advanced Cleaning with Auxiliary Features

The second preprocessing pipeline extends Methodology 1 by performing the aim of this approach is to investigate whether sentiment classification results can be improved and whether hand-crafted features can support transformer-based representations.

Emoji and non-ASCII characters are removed as follows:

$$T_4 = \{w_i \in T_3 \mid \text{ASCII}(w_i) = \text{true}\}$$

This step reduces the noise associated with encoding and maintains tokenization process compatibility. The token sequence is then adjusted to a fixed maximum length $L_{\max}$:

$$T_6 = \begin{cases} T_5[1:L_{\max}], & |T_5| > L_{\max} \\ \text{pad}(T_5), & |T_5| \leq L_{\max} \end{cases}$$

This guarantees the same input size of the transformer model.In addition to the processed text, three additional features are extracted:

$$\text{TokenLength} = |T_6|$$
$$\text{ExclamationCount} = \sum I(w_i = \text{"!"})$$
$$\text{QuestionMark} = \begin{cases} 1, & \text{"?"} \in T_6 \\ 0, & \text{otherwise} \end{cases}$$

Structures such as emphasis and questioning patterns, which may be associated with emotions, are captured in these features. To maintain consistency, the same approach as Methodology 1 is used for sentiment labelling and class balancing. The complete preprocessing workflow is shown in Fig. 3.

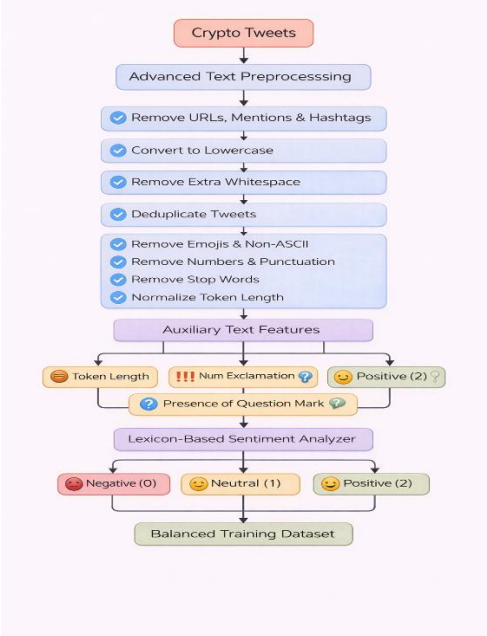

Fig. 3. Advanced preprocessing pipeline with extra feature extraction used in Methodology 2.

F. Train–Test Split Strategy

To evaluate the model performance and generalization, the processed and balanced datasets were tested using three train-test split ratios: 60% training and 40% testing, 70% training and 30% testing, and 80% training and 20% testing. To maintain a similar class distribution in the training and testing

sets, class-wise sampling was used. By using multiple split configurations, model performance can be compared under different training data sizes and the dependence on a single data partition can be reduced.

6. TRANSFORMER MODEL TRAINING

The cleaned and balanced cryptocurrency tweet dataset was used to further train three pretrained transformer models: RoBERTa, DeBERTa, and FinBERT. RoBERTa was chosen because it performs well on general emotion-related tasks and can also handle informal or noisy text well. DeBERTa was chosen because of its attention structure and representation method, which improves context understanding. FinBERT was included because it is specifically designed for financial texts. The use of these three models allows the system to identify different language patterns in cryptocurrency tweets.

Each tweet contains tokens that can be represented as follows:

$$X = \{x_1, x_2, \dots, x_n\},$$

where $n \leq L_{max}$ and $L_{max} = 64$ is the maximum sequence length. Maintaining a fixed sequence length helps manage memory usage and ensures consistent input sizes across batches. This approach preserves enough context for short tweets, making training more stable over inputs of varying lengths. The input sequence is converted into a contextual representation by a transformer encoder:

$$H = \text{TransformerEncoder}(X),$$

Where $H \in \mathbb{R}^{n \times d}$ and d represent the size of the hidden dimension. Each token can consider all other tokens in the sequence due to self-attention. In this way, long-range relationships, changes in sentiment, and contextual signals that are difficult to model with recurrent or convolutional networks can be captured. Due to the informal and context-heavy nature of Twitter text, Transformer encoders are ideal for this task.

The Special Classification Token ([CLS]) representations are extracted for multi-class sentiment classification, then passed through a linear layer and a softmax function:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b),$$

The predicted probabilities for the sentiment classes are denoted by $\hat{y}$, and the parameters W and b can be trained.

The problem is a multi-class task, so Softmax is used. Each tweet belongs to only one class: negative, neutral, or positive. Softmax raw model outputs into probabilities that sum to one. Other options, such as argmax, only return the final class label without any probability information. SparseMax can create sparse probabilities, but this is less common in this setting and may impact training stability. The multi-label problems benefit from sigmoid activation when multiple classes can be applied simultaneously. For these reasons, softmax is used in together with cross-entropy loss.

Loss Function

The model is trained by minimizing the categorical cross-entropy loss:

$$L = -\sum_{c=3}^{C} y_c \log(\hat{y}_c),$$

where $C = 3$ is the number of sentiment classes, y_c is the true label (one-hot encoded), and $\hat{y}_c$ is the predicted probability for class c. Because it directly measures the difference between the true and predicted probability distributions, cross-entropy is considered superior to the mean square error in improving classification.

Optimization

Model parameters are updated using the AdamW optimizer:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon} - \lambda \theta_t,$$

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05752
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

where η denotes the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected gradient estimates, λ is the weight decay coefficient, and ϵ is a small constant for numerical stability. When training large pretrained models, AdamW improves regularization by separating weight decay from gradient updates. All models were trained for six epochs using an NVIDIA GPU (CUDA) to speed up computation. To speed up the calculations, all models were trained with NVIDIA GPU (CUDA) for six epochs. The same training settings—learning rate, batch size, sequence length, and optimizer—were used for all models to ensure fair comparisons as shown in Table 2.

Table 2: Hyperparameter Configuration Table

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Sequence Length	64
Epochs	6
Loss Function	Cross-Entropy
Device	NVIDIA GPU (CUDA)

7. SYSTEM ARCHITECTURE OVERVIEW

Data preparation, model training, ensemble learning, and performance evaluation are all part of the step-by-step process that cryptocurrency sentiment analysis systems follow. The system uses a labelled offline Twitter dataset to classify sentiments into three categories: negative, neutral, and positive. After preprocessing and balancing, the tweets are divided using different train-test splits. RoBERTa, DeBERTa, and FinBERT process the cleaned text separately. A softmax layer is used to generate class probabilities for each model.

The weighted soft-voting technique combines the outputs of three models to increase prediction consistency. This strategy reduces prediction errors and makes the final output more stable. The Accuracy, precision, recall, F1-score, specificity and ROC-AUC of different train-test splits are used to evaluate the performance.

A. System Architecture Block Diagram

The workflow of the entire system is shown in Fig. 4. It outlines the process from the initial data collection to the final ensemble-based sentiment evaluation.

Fig. 4. System Architecture of the Proposed Cryptocurrency Twitter Sentiment Analysis Framework

8. PERFORMANCE EVALUATION METRICS AND RESULTS

A. Performance Evaluation Metrics

Accuracy, precision, recall, F1-score, and specificity are common multi-class classification metrics that were used to evaluate the system. Macro-averaging was applied to give equal weight to each sentiment class, regardless of how many samples belonged to each class.

In addition to these metrics, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were also used to measure how well the models differentiate between classes. The ROC curve shows the relationship between the true positive rate (TPR) and the false positive rate (FPR). Higher AUC values indicate better class separation.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05752
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Metric Definitions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where
TP = True Positives,

TN = True Negatives, FP = False Positives,

FN = False Negatives.

The combination of these metrics clearly shows the accuracy and error behaviour of the mode.

B. Performance Comparison of Individual Transformer Models

As shown in Tables 3–5, A comprehensive evaluation of individual Transformer models. RoBERTa, DeBERTa, and FinBERT – in various train-test split configurations and macro-averaging was used to calculate the F1-score, specificity, accuracy, precision, and recall for each model:

Table 3: RoBERTa Performance

Metric	60–40	70–30	80–20
Accuracy	82.37%	82.22%	82.69%
Precision	82.21%	82.09%	82.92%
Recall	82.15%	81.96%	82.86%
F1-Score	80.18%	82.02%	82.76%
Specificity	90.10%	91.33%	91.42%

Table 4: DeBERTa Performance

Metric	60–40	70–30	80–20
Accuracy	81.71%	83.61%	83.08%
Precision	81.16%	82.81%	83.80%
Recall	81.89%	82.64%	83.73%
F1-Score	82.00%	83.50%	83.09%
Specificity	90.85%	91.81%	91.54%

Table 5: FinBERT Performance

Metric	60–40	70–30	80–20
Accuracy	80.08%	80.11%	81.33%
Precision	80.59%	80.68%	81.87%
Recall	80.42%	79.51%	80.71%
F1-Score	80.02%	80.90%	81.56%
Specificity	90.04%	90.06%	90.67%

Observation:

The results illustrated in Fig. 5 and summarized in Table 3-5 shows that performance improves as the amount of training data increases. Under the 80–20 split, DeBERTa achieves the highest scores across most evaluation measures. RoBERTa maintains steady and competitive performance across splits. FinBERT performs well when it is fine-tuned on the domain-specific dataset

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05752

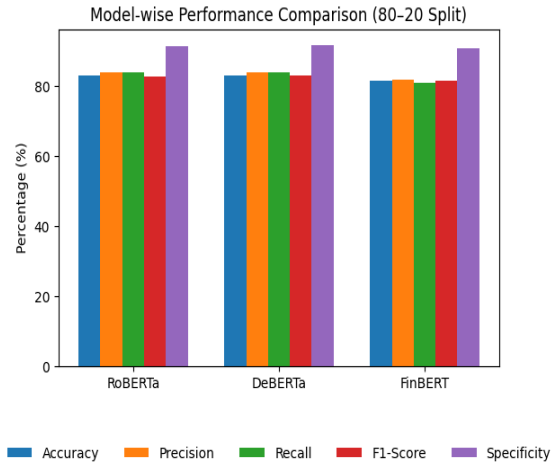

Fig. 5. Metric-wise comparison of individual transformer models across different train–test splits

C. Ensemble Model Performance

A weighted soft-voting ensemble was created by combining RoBERTa, DeBERTa, and FinBERT to increase prediction stability and reduce prediction variance. DeBERTa was given more weight in the group because it performed better than the other two models in different splits. The Performance of ensemble model under each split is summarized in below Table 6 and illustrated in Fig. 6.

Table 6: Ensemble Model Performance Across various Train–Test Splits

Train–Test Split	Accuracy	Precision	Recall	F1-Score	Specificity
60–40	84.46 %	83.39 %	83.28 %	84.47 %	90.67%
70–30	84.33 %	83.47 %	83.41 %	83.31 %	92.17%
80–20	86.00 %	85.86 %	85.79 %	85.82 %	92.46%

Observation:

In every split, the ensemble model outperforms all individual models. The best performance is observed under the 80–20 split. This shows that weighted soft voting improves the overall classification reliability and consistency.

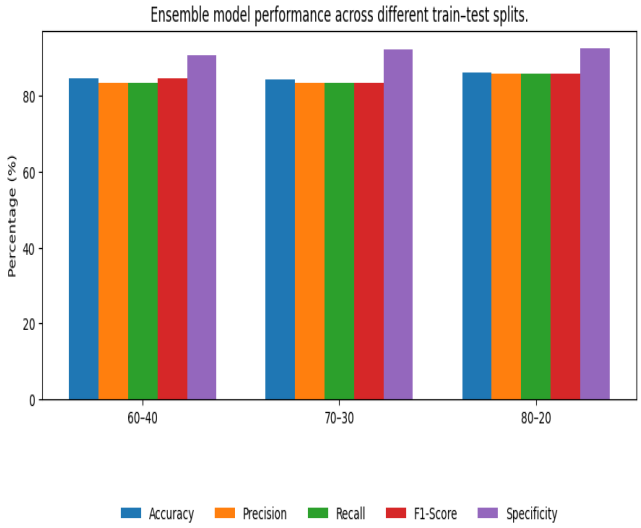

Fig. 6. Bar Plot illustrating Ensemble model performance across different train–test splits.

D. ROC Curve and AUC Analysis

To test how well the ensemble separates the three sentiment classes, ROC curves and AUC scores were calculated using the one-versus-rest (OvR) approach. The relationship between the true positive rate (TPR) and the false positive rate (FPR) is shown by the ROC curve in Fig. 7. AUC provides a performance measure that does not depend on any specific classification threshold. All splits show that the ensemble has high AUC values, as indicated by the results. The 80-20 split has the highest AUC, which indicates significant class separation.

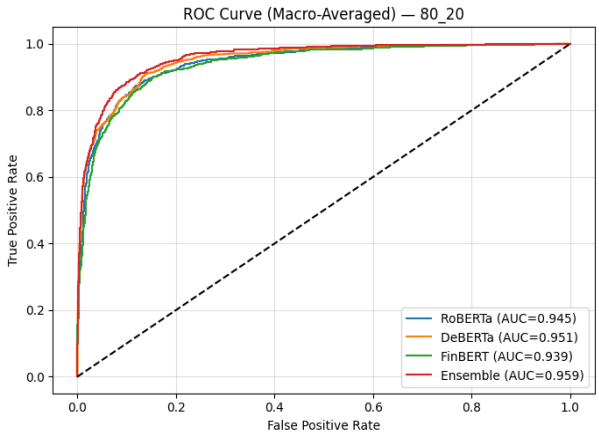

Fig. 7. Multi-class ROC curves of the Individual and ensemble model under the 80–20 train–test split.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM05752
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

D. Final Comparative Analysis

The 80–20 train-test split, which produced the best overall experimentation results, was used to directly compare the relative performance of each architecture. The table below summarizes the evaluation metrics for RoBERTa, DeBERTa, FinBERT, and the weighted ensemble model. Table 7 shows the evaluation metrics for RoBERTa, DeBERTa, FinBERT, and weighted ensemble models. Fig 8 shows the comparative performance of individual transformer models and ensembles under the same 80–20 split configuration.

Table 7: Comparative Performance of Models Under the 80-20 Train-Test Split

Metric	Accuracy	Precision	Recall	F1-Score	Specificity
RoBERTa	82.69 %	82.92 %	82.86%	82.76%	91.42 %
DeBERTa	83.08 %	83.80 %	83.73%	83.09%	91.54 %
FinBERT	81.33 %	81.87 %	80.71%	81.56%	90.67 %
Ensemble Model	86.00 %	85.86 %	85.79%	85.82%	92.46 %

Key Findings:

- The ensemble outperforms all the individual models.
- Weighted soft voting improves stability, specificity, and error balance.

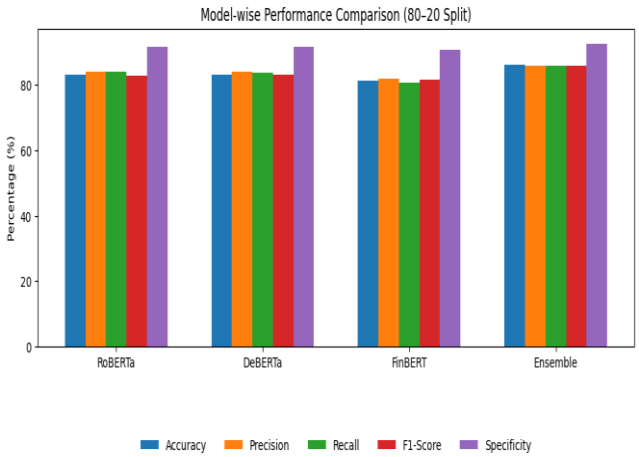

Fig. 8. Comparative analysis of individual transformer models and the ensemble model under the 80–20 split.

9. DISCUSSION

The results show that transformer-based models are effective for multi-class sentiment classification of cryptocurrency tweets. As the training data grew, it steadily improved, and the highest scores were achieved in an 80-20 split. This confirms that transformer architectures can stabilize contextual learning with sufficient fine-tuning data [3] [11]. Under the 80-20 configuration, DeBERTa performed the best among the individual models (83.09% F1-score) due to its convolutional attention mechanism, which improves contextual discrimination [13]. Strong generalization ability was demonstrated by the stable performance of RoBERTa across splits (82.76 % F1-score) [12]. FinBERT competitive results (81.56% F1 score) required domain-specific fine-tuning, indicating that financial-domain pretraining alone is insufficient for informal cryptocurrency text [5][14].In 80-20 split, the weighted soft-voting ensemble performed best overall, with an F1 score of 85.82% and an accuracy of 86.00%. This is a +2.92% improvement in accuracy compared to RoBERTa and DeBERTa and a +2.73% gain in F1-score compared to the best standalone model. Moreover, Ensemble achieved a specificity of 92.46%, which shows better class balance. Consistent with previous ensemble

studies [6] [16], these findings confirm that integrating multiple transformer architectures improves prediction reliability and reduces prediction variance. Transformer-based integration provides advanced contextual modelling and measurable performance enhancements compared to recurrent-based cryptocurrency ensemble [7].

This work evaluates three configurations – 60-40, 70-30 and 80-20 – unlike studies that rely on a single train-test split. Through improved generalization stability and reduced sensitivity to data partitioning, the ensemble consistently outperformed the individual models across all splits. These findings are further supported by the ROC-AUC analysis, which shows that the class separation is robust and the AUC values are stable and high. The preprocessing strategy also contributed to the model's effectiveness. Abbreviations, emojis, and informal expressions are common in cryptocurrency tweets, which can introduce noise [2]. The Controlled text normalization reduced irrelevant patterns while retaining sentiment-generating words. By combining lexicon-based labelling with contextual transformer learning, continuous supervision without manual annotation became possible.

In summary, the findings highlight three important points:

1. Transformer models are perfect for cryptocurrency sentiment classification when fine-tuned properly.
2. Performance in contextual learning and classification is improved by increasing the training dataset.
3. Compared to individual models, a weighted group of multiple transformer architectures provides greater stability and predictive balance.

These results extend previous studies on Transformer-based sentiment analysis [11–13] and ensemble learning [6, 7] in financial text classification. A more accurate assessment of the running performance of models in dynamic social media environments can be done by evaluating them across multiple train-test splits.

10. LIMITATIONS

Despite the framework's impressive results, it has some drawbacks. The evaluation was conducted using a previously collected static dataset. The volatile nature of cryptocurrency discussions on social media may not be fully reflected in this. The behaviour of the system has not been tested under constantly changing market trends. The current framework does not support real-time Twitter streaming. Therefore, it cannot be used for live sentiment tracking or rapid market changes. Instead of using automated optimization methods, ensemble weight specifications were determined based on observed trips. Transformer models also require a lot of computing power for inference and training. When combined, this increases computational costs, which may limit use in real-time deployments or resource-limited environments.

11. CONCLUSION

This study presented a sentiment analysis framework for cryptocurrency-related Twitter data using transformer models and ensemble learning. The Cleaning, normalizing, labelling, and balancing noisy social media text while retaining essential sentiment information was performed using two structured preprocessing methods. This ensured consistent input quality and unbiased model evaluation. Three Transformer models—RoBERTa, DeBERTa, and FinBERT—were tested in different train-test splits (60-40, 70-30, and 80-20). Although different models were successful trained, their stability varied with the size of the training set. Overall, with weighted soft voting, the ensemble consistently outperformed the individual models in terms of accuracy, F1-score, and overall reliability. When ensemble learning and transformer representations are combined, both stability and prediction variance are reduced. This framework can support research-oriented decision analysis in financial markets and provides a structured approach to cryptocurrency sentiment analysis.

Conflict for interest : Nil

REFERENCES

1. Yang, S.Y., Mo, S.Y.K., Liu, A.: Twitter financial community sentiment and its predictive relationship to stock market movement. Quant. Finance. 15, 1637–1656 (2015). https://doi.org/10.1080/14697688.2015.1071078.
2. Ramtin Ardeshirifar: Comparing hand-crafted and deep learning approaches for detecting AI-generated text: performance, generalization, and linguistic insights. (2019). https://doi.org/https://doi.org/10.1007/s43681-025-00699-4.
3. Huang, Y., Xu, J., Lai, J., Jiang, Z., Chen, T., Li, Z., Yao, Y., Ma, X., Yang, L., Chen, H., Li, S., Zhao, P.: Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey. (2024).
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).
5. Araci, D.: FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. (2019).
6. Lin, S., Zheng, H., Han, B., Li, Y., Han, C., Li, W.: Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. Acta Geotech. 17, 1477–1502 (2022). https://doi.org/10.1007/s11440-021-01440-1.
7. Aslam, N., Rustam, F., Lee, E., Washington, P.B., Ashraf, I.: Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model. IEEE Access. 10, 39313–39324 (2022). https://doi.org/10.1109/ACCESS.2022.3165621.
8. Dadhich, A., Thankachan, B.: Opinion Classification of Product Reviews Using Naïve Bayes, Logistic Regression and Sentiwordnet: Challenges and Survey. IOP Conf. Ser. Mater. Sci. Eng. 1099, 012071 (2021). https://doi.org/10.1088/1757-899x/1099/1/012071.
9. Jatnikaa, D., Bijaksanaa, M.A., Suryania, A.A.: Word2Vec Model Analysis for Semantic Similarities in English Words. ScienceDirect. 157, (2019). https://doi.org/10.1016/j.procs.2019.08.153.
10. Hai Nguyen Kiyoaki Shirai, T.: Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction.
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North. pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423.
12. Semary, N.A., Ahmed, W., Amin, K., Pławiak, P., Hammad, M.: Improving sentiment classification using a RoBERTa-based hybrid model. Front. Hum. Neurosci. 17, (2023). https://doi.org/10.3389/fnhum.2023.1292010.
13. Liu, J., Zhang, Z., Lu, X.: Aspect Sentiment Classification via Local Context-Focused Syntax Based on DeBERTa. In: 2024 4th International Conference on Computer Communication and Artificial Intelligence, CCAI 2024. pp. 297–302.

Institute of Electrical and Electronics Engineers Inc. (2024). https://doi.org/10.1109/CCAI61966.2024.10603339.

14. Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.: FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. (2020).

15. Khedr, A.M., Arif, I., Pravija Raj, P. V., El-Bannany, M., Alhashmi, S.M., Sreedharan, M.: Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey, (2021). https://doi.org/10.1002/isaf.1488.

16. Su, Y., Zhang, Y., Ji, D., Wang, Y., Wu, H.: Ensemble learning for sentiment classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 84–93 (2013).https://doi.org/10.1007/978-3-642-36337-5_10.

17. A. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011, doi: 10.1016/j.jocs.2010.12.007.

18. K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 2823–2824, doi: 10.1109/BigData.2015.7364089.

19. Y. M. Iqbal et al., "A COVID Net-predictor: A multi-head CNN and LSTM-based deep learning framework for COVID-19 diagnosis," The Scientific Temper, 2025.

20. Y. M. Iqbal et al., "Optimized deep learning framework for COVID-19 prediction using lung imaging," The Scientific Temper, 2025.

21. A. Saleem Raja, S. Peerbasha, Y. Mohammed Iqbal, B. Sundarvadivazhagan, and M. Mohamed Surputheen, "Structural Analysis of URL For Malicious URL Detection Using Machine Learning", *JOAASR*, vol. 5, no. 4, pp. 28–41, Jul. 2023.

BIOGRAPHIES

Dr. Y. Mohammed Iqbal is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over eight years of teaching and research experience. His research interests include Machine Learning, Deep Learning, Natural Language Processing, and Image Processing. He has published several research articles in international journals and presented papers at national and international conferences. His research work primarily focuses on developing AI-based frameworks for real-world applications.

Mohamed Fawaz S is a Master of Computer Applications (MCA) student at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. His areas of interest include Full-Stack Web Development, Machine Learning, and AI-integrated application development. He has experience in developing intelligent web systems using the MERN stack and has worked on projects involving transformer-based sentiment analysis for cryptocurrency tweets. He has also developed scalable web platforms and AI-assisted applications during his academic projects and industry internships.

Dr. S. Peerbasha is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and over Seventeen years of teaching and research experience. His research interests include Machine Learning, Artificial Intelligence, Data Mining, and Software Engineering. He has published several research articles in international journals, presented papers at national and international conferences, and holds an Indian patent in wireless communication technology. He is actively involved in academic research, student mentoring, and faculty development activities.

Dr. M. Mohamed Surputheen is an Associate Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 34 years of teaching and research experience. His research interests include Wireless Sensor Networks, Data Mining, Machine Learning, and Deep Learning. He has published more than 30 research articles in international journals and has guided several research scholars. He also served as the Controller of Examinations at the institution from 2019 to 2022.

Dr. M. Rajakumar is an Associate Professor and Research Advisor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 20 years of teaching and research experience. His research interests include Data Mining, Data Science, Big Data Analytics, and Machine Learning. He has supervised several M.Phil. and Ph.D. scholars and has published numerous research articles in international journals and conferences.