The Nucleus **61**, No. 1 (2024) 1-9

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)
ISSN 2306-6539 (Online)

Emotion Detection for Cryptocurrency Tweets Using Machine Learning Algorithms

Bushra Fareed[1], Mujeeb Ur Rehman[2*], Mumtaz Ali Shah[2], Akbar Hussain[2], Khudija Bibi[3]

[1]*Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, Pakistan*

[2]*University of Management and Technology, Sialkot, Pakistan*

[3]*International Islamic University, Islamabad, Pakistan*

ABSTRACT

Cryptocurrencies, functioning as digital currencies, undergo regular fluctuations in the present market, reflecting the emotional aspect of the cryptocurrency realm. It is a well-established fact that sentiment is linked to Bitcoin and Ethereum values, employing a Twitter-based strategy to predict changes. While prospective Bitcoin returns do not display a correlation with emotional variables, indicators of emotions tend to anticipate Bitcoin exchange volume and return volatility. Emotions wield an influence over a broad spectrum of financial investor returns, thereby, potentially affecting market dynamics by triggering significant price shifts. The research delves into gauging emotional factors extracted from 2,050,202 posts on Bitcointalk.org, investigating how these emotions impact Bitcoin's price fluctuations. We have used a unified dataset named 'data F' in which all categories of emotions are consolidated. Subsequently, data preprocessing steps are implemented to cleanse the dataset. Two feature engineering techniques, namely TF-IDF and BoW are employed. The research explores ten supervised machine learning (ML) models as classifiers, with four of these models (LR, Stochastic Gradient Descent, SVM and GB) yielding the highest accuracy at 0.93%.

Keywords: *Emotion, Twitter, Bitcoin, Ethereum, Cryptocurrencies*

1. Introduction

Over the course of an extensive period, artificial intelligence (AI) has evolved from a concept relegated to the realm of fiction to a practical force in the real world, thanks to the availability of adequate computing power for execution. Cryptography plays a pivotal role in safeguarding cryptocurrencies and their transactions [1]. Unlike traditional currencies reliant on central financial institutions, cryptocurrencies operate on the principle of decentralized control. Consequently, cryptocurrencies enable electronic money transfers without the need for intermediaries or conventional financial establishments. With their inherent characteristics of being uncontrollable and untraceable, the cryptocurrency industry has witnessed rapid growth within a short span of time. Virtual currencies are progressively being integrated into economic transactions across various domains. This field has garnered significant attention and a multitude of scholars are scrutinizing cryptocurrencies for diverse purposes, including financial prognostications and more. Researchers have displayed escalating interest in the commercial applications of cryptocurrencies. However, the utility of cryptocurrency and its associated technologies extends beyond the financial spheres. Numerous information technology (IT) courses encompass cryptocurrency technologies that hold the potential to devise novel and efficient methodologies for managing bitcoin, other cryptocurrencies, their price volatility, and related technologies. This article provides an overarching view of cryptocurrency price prediction research spanning the years from 2010 to 2020 [16]. It encompasses seminal studies concerning the prediction of cryptocurrency prices, incorporating both ML methods and statistical approaches. Furthermore, this study delves into datasets, trends, research methodologies and forecasting techniques, subsequently exploring uncharted territories within the realm of bitcoin price prediction research.

From May 2018 onwards, the two most prominent cryptocurrencies, as measured by market capitalization, held a combined total value of approximately $160.9 billion [2]. Out of this, bitcoin alone accounted for $115 billion of the total value. These digital currencies are believed to have value, functioning as real currencies and serving as investment opportunities for certain investors. Both of these currencies experienced significant fluctuations in value within a short time span. In the midst of 2017, the value of a single bitcoin surged by an astonishing two thousand percent, soaring from $863 on January 9 to $17,550 on December 11 of the same year. Merely eight weeks later, by February 5, 2018, the price of a single bitcoin had exceeded $79,643, again marking a two thousand percent increase [12]. The input for the ML models, specifically decision tree (DT), logistic regression (LR), and random forest (RF), involves utilizing user reviews about cryptocurrencies. The objective is to predict the emotional tones of these reviews, namely: joy, sadness, hatred, fear, anger or greed. The dataset employed in this research was sourced from Kaggle and it underwent preprocessing using the natural language toolkit (NLTK) in Python. NLTK was chosen due to its effectiveness in handling human language data for statistical natural language processing tasks. It facilitated the removal of stop words and other text cleaning procedures. Additionally, the text was converted to lowercase and underwent further preprocessing steps including tokenization, stop word removal, porter stemming and more. Upon completing the preprocessing phase, text feature extraction techniques such as term frequency and inverse document frequency were applied. The bag of words (BoW) approach was utilized to identify the most relevant features. Subsequently, the dataset was divided

[*]Corresponding author: mujeeb.rehman.pak@gmail.com

into two segments: a training set (80%) and a testing set (20%). To conduct the classification of reviews into emotional categories (joy, sadness, fear, hatred, anger and greed), a range of classification algorithms were employed. These included random forest (RF), decision tree (DT), stochastic gradient descent classifier, K-nearest neighbors (KNN), Ada boost classifier (AC), Gaussian Naive Bayes (GNB) and extra tree classifier (ETC). Notably, the first seven models are based on tree ensemble methods, while logistic regression (LR) was used as a distinct classification approach.

The fundamental aspects of employing this approach encompass:

- Categorization of data into six emotional classes. This aids individuals involved in cryptocurrency in comprehending people's emotions, thus, aiding their decision-making. Diverse ML models were chosen for this study, each with distinct parameters. These parameters were fine-tuned through empirical methods to attain optimal levels of accuracy, precision, recall and F1-score.
- Data can be segmented into six emotional types: joy, sadness, fear, hate, as well as greed and anger. This categorization is based on user reviews.
- Application of preprocessing techniques, including converting to lowercase, replacing characters from A to Z, splitting, eliminating stop words, applying porter stemmer, cleansing tweet reviews and utilizing effective learning models.
- Implementation of techniques like: term frequency and inverse document frequency (TF-IDF) and BoW through feature engineering.

2. Related Work

This section provides a concise evaluation of cryptocurrency price prediction methods. The related research can be categorized into three main types: (1) assessment of social media influence on financial markets, including cryptocurrency markets; (2) employment of ML for predicting cryptocurrency prices; and (3) utilization of extensive data frameworks for financial market prediction.

Daniel et al. investigate decision-making growth and risk through alternate models in prospect theory [2, 3]. Typically seen in financial psychology, these models depict a range of sentiments that substantially impact a financial agent's decision-making process, consequently leading to a consistent pattern of price projection [4]. These insights thus pave the way for employing methods like sentiment analysis to identify patterns affecting asset prices.

Regarding media's evolution, particularly on social platforms and their impact on user sentiment in financial sectors, Tetlock's study [5] distinguishes high negativity on social platforms as indicative of downward market pressure, with abnormally high or low negativity predicting high trading volume. Moreover, the majority of consumers resort to social networks for purchase decisions, yielding investigations that explore connections between sentiment in media (such as reviews) and various financial sectors. Heaton's research [6] retrieves, extracts and examines media's effects on financial markets. By constructing a sentiment analysis lexicon specific to the financial domain, the study achieved 70.59 percent accuracy in short-term stock market trend prediction. Stock performance prediction also extends to comment board analysis [7] uncovering the 'topic sentiment' feature reflecting emotions tied to industry-specific subjects and stock projections. This approach has achieved a 2.07 percentage point performance gain over one year with 18 trades, solely using historical prices. Similarly, Alan et al. linked tweet sentiment scores to stock prices [8], achieving an 86.7 percent accuracy in Dow Jones industrial average prediction using a self-constructed fuzzy classifier network. The ensuing increase in cryptocurrency speculation led to efforts to forecast price fluctuations [9]. Jafar et al. achieved successful predictions of price variations for Ethereum, Bitcoin, and Litecoin [10] by leveraging news and social media data annotated with actual prices. The method precisely anticipated price changes, achieving 43.9 percentage points accuracy for price increases and 61.9 percentage points accuracy for price decreases in bitcoin forecasts. Additionally, Twitter sentiment and Google trends data were used for bitcoin and ethereum price prediction [2]. The approach incorporated tweet volume and established a correlation with cryptocurrency prices. Nomiizz further expanded on this, utilizing sentiment analysis of tweets to establish a connection with bitcoin prices [11, 12]. Tweets underwent preprocessing to remove non-alphanumeric characters and were analyzed using the valence aware dictionary and sentiment reasoner (VADER) to categorize them as positive, neutral, or negative. Compound sentiment scores were then correlated with bitcoin prices across different time intervals. Nomiizz's approach extends our previous discussions, presenting a refined forecasting model for bitcoin prices over specified intervals.

Additionally, ML is a pivotal tool for cryptocurrency price forecasting. In a notable instance, the authors of [13] contrast bitcoin price prediction methods by benchmarking traditional auto-regressive integrated moving average (ARIMA) and ML-based neural network auto-regressive (NNAR) models against historical price data. Similarly, Lucey et al. [14] introduce classifiers employing a hierarchy of deep ML models for cryptocurrency price forecasting. This hierarchy includes a multi-layer perceptron (MLP), a basic recurrent neural network (RNN) and a long short-term memory (LSTM) network, specifically designed for complex sequence prediction. Notably, this study expands prediction methodologies by incorporating the impact of social media and applying online learning for better management. This approach accommodates an indefinite timeline for data generation. Consequently, our treatment of the problem aligns with a big data perspective, utilizing substantial data tools to ensure scalability and performance. We draw inspiration from Munim et al. [15] who utilize Apache Spark to analyze market trends on the foundation of social network data and historical prices. Naila Aslam and her team engage in cryptocurrency tweet detection and sentiment analysis, employing five emotion classes - happy, sad, fear, angry and surprise - utilizing techniques like BoW, TFIDF and Word2Vec [16]. Table 1 presents an extensive summary of various models utilized on a range of sentiment reviews gathered from diverse sources.

Table 1. Summary of Related Work

Ref. No.	Year	Models	Focus	Restriction	Dataset Size	Sources
2	2013	Descriptive model	Prospect Theory	No ML models.	N/A	N/A
3	2017	Efficient Market Hypothesis (EMH)	Feelings affect Investor Decision-Making	No ML models.	N/A	psychologists and behavioral economists
4	2007	Linear combination	Giving Content to Investor Sentiment	This article only discusses interactions between media content and stock market activity.	77 × 77	Dow Jones Newswires, Wall Street Journal column daily (1889 to 1897)
5	2017	Kaiser-Meyer-Olkin (KMO) and Bartlett's Test	Role of Google and Social Networks on Consumers' Buying Behavior	This article only explains 61.53% variety.	160	Google and Social Networks
6	2018	Hybrid Model (ARIMA & LSTM)	Predicting the Effects of News Sentiments on the Stock Market	This article only using news sentiments achieved only 70.59% accuracy in the short-term movement of stock price.	6 Months	specific news articles
7	2017	Support Vector Machine (SVM)	Sentiment Analysis on social media for Stock Movement Prediction	In this article, only a few specified numbers of topics and sentiments beforehand for Joint Sentiment Topic (JST)-based and the Latent Dirichlet Allocation (LDA).	1 Year	social media
8	2014	Dow Jones Industrial Average (DJIA)	Trading on Twitter	No ML models.	3,475,428	Tweets
9	2005	DJIA & Self-Organizing Fuzzy Neural Network	Role of Twitter Mood on the Stock Market	This study did not use ML models.	June 2009 to December 2009	Tweets
10	2017	LR, Bernoulli Naive Bayes & NB	Cryptocurrency Price Prediction Using News and Social Media Sentiment	This article finds the percentage of 67 days increases and decreases the price of bitcoin & Ethereum.	67 Days	Crypto Market
11	2018	Sentiment analysis	Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis	No ML models.	30,420,063	tweets and Google Trends data
12	2007	Sentiment analysis	Role of Twitter Sentiment Analysis in Bitcoin Price Fluctuation	No ML models.	N/A	Tweets
13	2021	VADER, vaderSentiment, Polarity classification & lexicon-based approach	Drabble/TwitterSentimentAndCryptocurrencies	Miss-tuning of hyperparameters	828'338	Tweets
14	2019	ARIMA & NNAR	Next-Day Bitcoin Price Forecast	This article only uses two test sample forecast periods.	Two Sample	Crypto Market
15	2018	MLP& RNN, LSTM	Deep Neural Network for Cryptocurrencies Price Prediction	In this article, use learns long dependencies.	1 Year	Crypto Market
16	2018	Hybrid model-based (Apache Spark & Hadoop HDFS)	Stock Market Real-Time Recommender Model Using Apache Spark Framework	In this article historical price is not used as the primary factor for a prediction of the stock market trend.	1-2-2013 to 30-6-2016	interval
17	2022	LSTM and GRU	Sentiment Analysis and Emotion Detection on Cryptocurrency	y Work five Classes	40,000	Tweets

Table 2 Sample Data

A	B	C
0	Crypto Panic: US Securities Watchdog Says Desperate Kik Turned to Blockchain Tokens https://t.co/CR3tmIuuZ8 Crypto https://t.co/hmEsmDINxh	Fear
1	Earning #cryptocurrency for selling my stuff on @Listia! Join me using code "DGWCDX" for an extra 100 XNK. I just l https://t.co/cYSFzxWZx5	Fear
3963	Let us stop annoying the big exchanges and screw ourselves over.. this is not helping anyone	Hate
8739	Happy Children's Day. Visit https://t.co/2OKeAHtIPH to buy a cryptocurrency and pay with Naira. https://t.co/hUItHTdnkt	Joy
12488	RT @CrptoKryptonite: @blockonix_com @Tronfoundation @Cardano @block_one_ @creditscom will be period. DYOR... low trans fees, most scalable	Sad
15051	Coinbase's crypto debit card arrives in 6 more countries #Cryptocurrencies #bitcoin #crypto #cryptocurrency https://t.co/x3E1Zmj8dt	Greed
21572	RT @jonhumbert: 1) Seditious traitor 2) Seditious traitor with a nickname 3) Total unit/hero 4) Angry boi/hero 5) Seditious traitor who was traitor	Anger

3. Methods and Material

In this study, different ML methods were used for this analysis. Various experiments were investigated using different methods and techniques, classification flowcharts, diagrams, evolutionary parameters and more. Several classifiers were evaluated for this purpose; however, the best classification was the voting classification.

The dataset was first downloaded from Kaggle (https://www.kaggle.com/datasets/huseinzol05/twitter-emotion-cryptocurrency). It was processed by removing irrelevant elements and divided into test and train datasets. The size of the test data set weighed 20% and the size of the training dataset is 80% weighed. The next step is the feature engineering kept in the training package. The test data set was forecast using classification. After the set of data was predicted, evolution parameters were applied to it to achieve faultless accuracy. The following parameters have been used to model evolution: precision, recall, F1-score, and accuracy.

The data for this experiment is taken from Kaggle, which has a large number of opposing tweets. The type of the data file is JavaScript Object Notation (JSON type of file is an open standard format use to share the data, it used human-readable text to save and share the data). The dataset contains six different types of files such as: joy.json, sadness.json, fear.json, hateful.json, anger.json and greed.json. The size of dataset files like joy.json has (54565) reviews, the size of sadness.json has (9765) reviews, the size of fear.json is (50053) reviews, the size of hateful.json is (10206) reviews, the size of anger.json is (3596) reviews and the last file size of greed.json is (200177) reviews. We append all these six files in one single file name dataF and the total size of the dataF file is (328362) reviews. We take target all these files like joy.json takes targets as Joy and sadness.json is sad and fear.json takes target fear, in hateful.json takes targets hate and anger.json takes targets anger and also takes from greed.json is greed reviews. The dataset used in this study is shown in the following Table 2.

Column A of Table 2 shows the number of rows and column B shows the tweets about the cryptocurrency and column C show the target values.

Table 3 shows the total number of data in each file.

Table 3. Target Reviews of Cryptocurrencies

Joy	Sad	Fear	Hate	Greed	Anger
54565	9765	50053	10206	200177	3596

3.1 *Engineering of Feature*

Engineering of features used to find efficient features from the dataset to use training the models or another way, the best feature selected from the original dataset [17] . Robertson [18] finishes that engineering features can raise the working of ML models. ‘‘refuse in refuse out’ is a mostly saying in ML. Another way, more instructive data can create desirable results. Engineering of feature help to increase the accuracy of the raw dataset.

3.1.1 *Term Frequency and Inverse Document Frequency (TF-IDF)*

Term frequency-inverse document frequency is used to find and select features in the dataset. Text analysis and music information retrieval use to term frequency-inverse document frequency [19]. The weight is assigned to every word in the document using TF-IDF [15, 20]. Any term or word that has a higher weight means more importance than other terms [21]. This formula is used to find every word or term weight as mentioned in Equation 1.

$$W_{i,j} = TF_{i,j}(\frac{N}{D_{f,t}}) \qquad (1)$$

In Equation 1 where N shows the number of documents and the $D_{f,t}$ shows the total number of words in the document.

3.1.2 *Bag of Words (BoW)*

A bag of words is a simple technique in which we convert simple text data into numeric form. On this numeric form data, we assign the frequency to each word in the data. Frequency depends on how many times the particular word comes in the corpus on the basis of word frequency we determine the importance of each word in sentimental analysis. This formula is used to find every word or term weight as mentioned in the equation.

3.2 Supervised Machine Learning Methods

In supervised ML, this study discusses the ML models and also shows the algorithms and their hyper parameters. ML uses to apply and implement algorithms by the Scikit-learn library and NLTK [22]. ML-supervised algorithms are mostly used for classification and regression [23]. We use RF, LR, DT, KNN, Ada Boost Classifier, STOCHASTIC GRADIENT DESCENT Classifier, ET, GNB and Gradient Boosting Classifier.

Table 4. ML Models and Hyper-parameters

ML models	Hyperparameters	Using
LR	Multi-class="multinomial"	In LR only work binary class that way we use Multi-class="multinomial" work for multiple classes
RF	n-estimators=50	Rang of n-estimators is 50 to 500 but we use only 50, which gives us the best results
DT	Random-state=0	For solving the problems for train and test split that way we use Random-state=0
KNN	n-neighbors=5	For the range of n-neighbors 1 to 5, we use the 5 that way we get the best results
SVM	kernel='linear'	is used when the data is linearly separable, that is, it can be separated by a single line

Ada Boost Classifier	n-estimators=50,
	learning-rate=1
		In the range of 0.0 to 1.0 learning rate we use 1 for the results

GNB	Priors=none,
	var_smoothing= 1 e-09
		The portion of the largest variance of all elements that is added to the variances for calculation stability

ExtraTree Classifier	Random-state=0	For solving the problems for train and test split that way we use Random-state=0
Stochastic Gradient Descent Classifier	Random-state=0	For solving the problems for train and test split that way we use Random-state=0
Gradient Boosting Classifier	Random-state=0	For solving the problems for train and test split that way we use Random-state=0

3.2.1 Random Forest (RF)

Random forests or random decision forests are an ensemble getting-to-know approach for classification, regression and other obligations that perform by using building a large number of selections at schooling time and outputting the class this is the mode of the training (classification) or suggesting/averaging prediction (regression) of the person branches.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2 \tag{2}$$

In Equation 2, where N is the wide variety of information factors, fi is the value returned with the aid of the version and yi is the actual cost for statistics point i.

3.2.2 Logistics Regression (LR)

Logistic regression is a statistical analysis technique used to expect an information price based totally on prior observations of an information set. A logistic regression version predicts a dependent statistics variable by using studying the connection among one or greater current independent variables.

3.2.3 Gaussian NB (GNB)

A Gaussian Naive Bayes set of rules is a unique kind of NB set of rules. It is in particular used while the functions have continuous values. It is also presumed that each one of the functions are following a Gaussian distribution, particularly the everyday distribution. A Gaussian classifier is a generative technique in the sense that it tries to version class posterior as well as enter class-conditional distribution. Consequently, we can generate new samples in the entering area with a Gaussian classifier.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

In Equation 3, A and B are two events where P(A|B) or P(B|A) is the probability of the events A and B. The independent probability P(A) of A and P(B) of B.

3.2.4 Extra Tree Classifier (ETC)

Extra Tree Classifier (ETC) is a gathering learning strategy on a very basic level based on decision trees. ETC like RF randomizes certain choices and subsets of information to play down over-learning from the information and over-fitting.

3.2.5 Decision Tree (DT)

The goal of employing a decision tree is to form an education model which can make use of to expect the lesson or esteem of the target variable by way of studying sincere choice policies deduced from in advance information (training facts). In choice bushes, for predicting a category name for a record we start from the root of the tree. We evaluate the values of the foundation property with the record's best. On the idea of comparison, we take after the branch comparing to that esteem and bounce to any other node.

$$E(S) = \sum_{i=1}^{C} -p_i p_i \tag{4}$$

In Equation 4, where S is a current state and the p_i is the probability of the ith state of S.

3.2.6 K. Neighbors (KNN)

K-nearest neighbor (KNN) can be an exceptionally basic, easy to get it, bendy and one of the topmost machine mastering algorithms. KNN is utilized in an assortment of applications inclusive of the fund, healthcare, political science, penmanship discovery, photograph acknowledgment and video acknowledgment. In credit score value determinations, financial organizing will anticipate the credit rating of customers. In credit price, keeping money set up will anticipate whether the credit is comfortable or risky. In political technological know-how, classifying capacity voters in classes will vote or receive the vote. KNN calculation is applied for each type and relapse issue. Moreover, it is based totally on the spotlight likeness approach.

3.2.7 Stochastic Gradient Descent (SGD) Classifier

Utilizing stochastic gradient descent (SGD) on regularized direct approaches can aid in constructing an estimator for both classification and regression problems. The Scikit-learn API provides a module to implement the strategy specifically for classification issues. The SGD classifier employs a regularized linear model with the aim to build an estimator. It demonstrates robust performance with large-scale datasets and is an efficient and straightforward strategy to implement.

$$\theta_j = \theta_j - \alpha\ (\widehat{y^i} - y^j)x_j^i \tag{5}$$

In Equation 5 where θ_j estimator of the $(\widehat{y^i} - y^j)x_j^i$

3.2.8 Support Vector Classifier (SVC)

The goal of a linear SVC is too healthy to the data, returning a "find match" hyperplane that categorizes statistics. Based on this hyperplane, you may then feed some functions for your classifier to look at what the "anticipated" magnificence is.

3.2.9 Ada Boost Classifier

The Ada Boost algorithm, short for adaptive boosting, serves as a boosting technique employed in ML. As an ensemble approach, it earns its designation as adaptive boosting by reallocating weights to each instance, assigning higher weights to inaccurately labeled occurrences.

Boosting is utilized to decrease inclination as well as the fluctuation for administered learning. It works on the rule where learners are developed consecutively. However, for the primary level, each ensuing learner is developed from already developed learners. In straightforward words, frail learners are changed over into strong ones. Ada Boost calculation too works on the same guideline as boosting, however, there is a slight contrast in working.

4. Proposed Methodology

This research employs a variety of techniques to address classification challenges. The approach for resolving the classification problem is depicted in Fig.1. The initial stage involves the processing of information during the initialization phase. The study categorizes this information into six classes: joy, sadness, fear, hate, greed and anger as shown in Table 5.

During the testing phase, assessment texts are converted to lowercase letters as a preliminary step. Following this, the process involves applying segmentation to critiques and subsequently utilizing the stemming method on the reviews. This is done to achieve the fundamental structure of each phrase. As part of the preprocessing stage, certain phrases that tend to create confusion in text analysis are removed from the textual content reviews. The outcomes of this preprocessing are illustrated in Table 6, displaying the sample data after undergoing these steps.

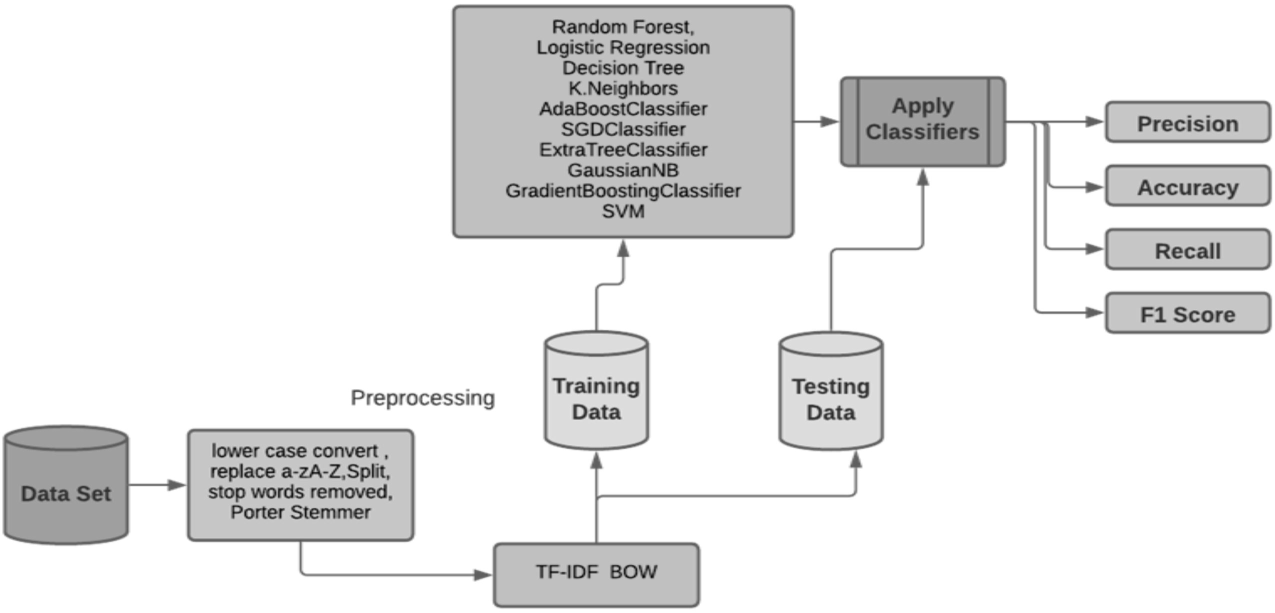

Fig.1 Proposed Methodology

Once the preprocessing is complete, the dataset is divided into subsets for both training and testing purposes. This division adheres to an 80-20 ratio, where 80% of the data is used for training and the remaining 20% is allocated for testing.

Table 5. Quantity of Samples corresponding to each target elegance

Tweets data	Target Value
I see the Tesla fan boys are angry at the idea	Anger
amritabithi You're gonna hate it when it is	Hate
This makes me very sad. Dwane was a great	Sad
The Fear and Greed Index big	Greed
The Fear and Greed Index big	Fear
+8.73% over the last hour JOY crypto	Joy

Table 6 Tweets Preprocessing

Without Preprocessing	With Preprocessing
I see the Tesla fanboys are angry at the idea	see tesla fanboy angry idea
amritabithi You're gonna hate it when it is	amritabithi go hate
This makes me very sad. Dwane was a great	Make very sad.d wane great
The Fear and Greed Index big	fear greed index big
The Fear and Greed Index big	fear greed index big
+8.73% over the last hour JOY crypto	+8.73%over last hour joy crypto

5. Evaluation

In this section, we engage in a discussion about solving grouping problems using various models. The utilization of TF-IDF and BoW techniques for feature selection is explored within the context of the article. The article employs TF-IDF and BoW methods, leading to more robust conclusions. A comparative analysis is tested empirically from models like RF, DT, SGD classifier, KNN, AC, GNB, GB classifier and ETC against the LR.

The time complexity of distinct ML models differs concerning both testing and training times. Although there's minimal variance in testing time among these models, training time displays significant fluctuations contingent on the complexity of the ML model. For instance, linear regression and Naive Bayes are simple models with linear time complexity, whereas SVM exhibits a complex nature with quadratic time complexity during the training phase. However, it's important to note that while our research focuses on accuracy and other pertinent parameters, the aspect of time complexity isn't directly intertwined with the primary contributions of our work.

Within our research, we deliberately select MLmodels with diverse parameters. These parameters are chosen through empirical methods to achieve optimal accuracy. The Gaussian Naive Bayes (GNB) classifier, operating on a likelihood basis, attains the lowest accuracy compared to other models. While KNN excels in accuracy for smaller datasets, its performance diminishes on our larger dataset. On the other hand, the amalgamation of linear models in random forest yields comparatively favorable results. Models such as ETC, DT and ADB exhibit enhanced accuracy, particularly for multi-class datasets, outperforming the GNB model. Notably, GB classifier, logistic regression, support vector machine (SVM) and SGD classifier exhibit the highest accuracy among the models employed in our research. This is due to the clarity and labeling present in our dataset, aligning with the conditions necessary for these models to excel. Conversely, should our dataset lack proper labeling or preprocessing, these models would perform sub optimally due to the potential noise present in the data. Given our dataset's probabilistic nature, the GNB model produces exceptional results across all models.

We empirically fine-tune these parameters to achieve peak accuracy. For instance, logistic regression, SGD classifier, SVM and GB classifier demonstrate superior accuracy, precision, recall and F1-score through the application of multinomial parameters. The outcomes of TF-IDF are presented in Table 7 to provide a comprehensive view of the results.

Table 7 Result in TF-IDF

Models	Accuracy	Precision	Recall	F1-score
RF	0.9	0.77	0.75	0.76
LR	0.93	0.84	0.8	0.82
GNB	0.36	0.36	0.53	0.33
ETC	0.87	0.72	0.72	0.72
DT	0.89	0.76	0.74	0.75
KNN	0.81	0.61	0.67	0.62
Stochastic Gradient Descent	0.93	0.86	0.79	0.84
SVM	0.93	0.84	0.8	0.81
Aboost	0.92	0.85	0.82	0.83
Gradient Boosting Classifier	0.93	0.80	0.80	0.81

In our investigation, the data utilized for this experimentation was sourced from Kaggle, a platform that hosts a substantial collection of opposing tweets. The data is stored in JSON format and consists of six distinct file types: joy.json, sadness.json, fear.json, hateful.json, anger.json and greed.json. While previous works on datasets lacked a diversity of classes conducive to aiding cryptocurrency miners in making more informed decisions, this study takes a different approach by categorizing data into six emotional classes. This classification empowers cryptocurrency minors to gain insights into people's emotions, thus enhancing their decision-making process.

Within this study, a variety of ML models with differing parameters were chosen. These parameters were meticulously fine-tuned through empirical methods to attain the highest possible levels of accuracy, precision, recall and F1-score. It's noteworthy that the GNB classifier yields the lowest values for accuracy, precision, recall and F1-score due to its likelihood-based operation, causing it to lag behind other models. Similarly, KNN performs exceptionally well with accuracy, precision, recall and F1-score in small datasets, but its performance diminishes when applied to our larger dataset.

The random forest (RF) model, which amalgamates various linear models, produces moderately improved outcomes. Conversely, models such as: ET, DT and ADB exhibit improved accuracy, precision, recall and F1-score due to their compatibility with multi-class datasets similar to ours.

However, it is the GB classifier, LR, SVM and SGD classifier that deliver the most impressive accuracy, precision, recall and F1-score among all models employed in our research. This superiority is attributed to our dataset's clear labeling, aligning perfectly with the conditions required for these models to excel. In contrast, if our dataset lacks proper labeling or preprocessing, these models perform inadequately due to potential noise, leading to subpar accuracy, precision, recall and F1-score.

Given our dataset's probabilistic nature, the Gaussian Naive Bayes model excels across all models. Fig. 2 visually represent applying GMB give less than others models accuracy, precision, recall and F1-score with TF-IDF. Figures 3 and 4 visually represent the contrast between the least and greatest values for accuracy, precision, recall and F1-score. Fig. 3 illustrates GNB's least favorable performance using TF-IDF, while also showcasing LR, SGD classifier, GB and SVM models' optimal performance. Fig. 4 shows comparative analysis for different selected models with BOW.

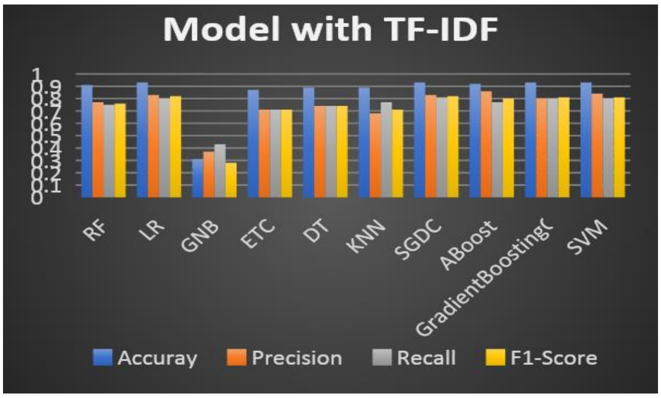

Fig. 2 Classifier Training-cross Validation Curve

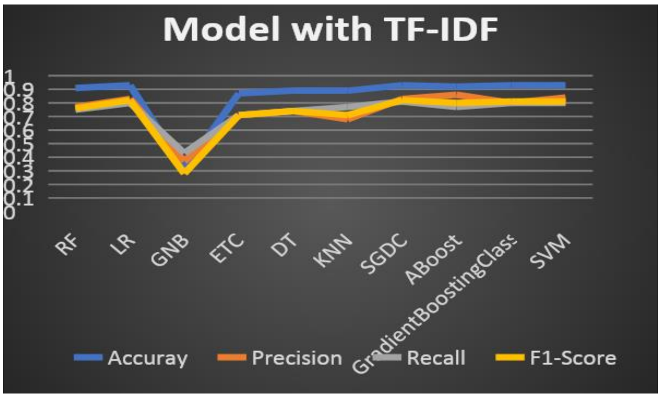

Fig. 3 Results of Models using TF-IDF

We determined these parameters empirically to attain optimal levels of accuracy, precision, recall and F1-score. As an illustration, the linear parameter yields the highest values for accuracy, precision, recall and F1-score among the SVM models. The outcomes for BoW are detailed in Table 8.

Table 8 Result in BoW

Models	Accuracy	Precision	Recall	F1-score
RF	0.91	0.77	0.75	0.76
LR	0.93	0.83	0.80	0.82
GNB	0.31	0.37	0.43	0.28
ETC	0.87	0.71	0.71	0.71
DT	0.89	0.74	0.74	0.74
KNN	0.89	0.68	0.77	0.71
Stochastic Gradient Descentc	0.93	0.83	0.81	0.82
Ada Boost	0.92	0.86	0.77	0.80
Gradient Boosting Classifier	0.93	0.80	0.80	0.81
SVM	0.93	0.84	0.8	0.81

Fig. 5 illustrates the minimum accuracy, precision, recall and F1-score achieved by GNB when utilizing BoW. Additionally, it displays the maximum accuracy, precision, recall and F1-score obtained by LR (Logistic Regression), SGD (Stochastic Gradient Descent), GB (Gradient Boosting) and SVM (Support Vector Machine).

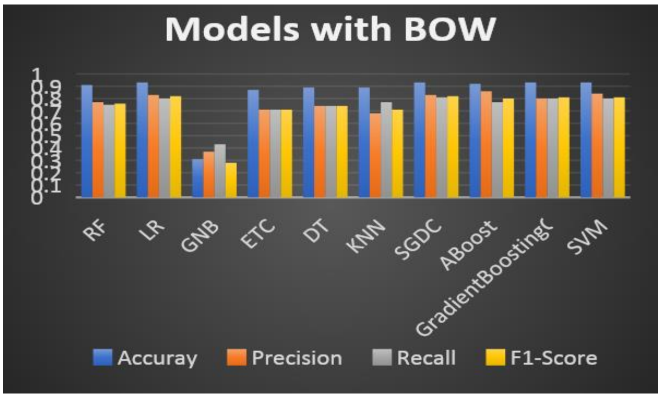

Fig. 4 Clustered Column Chart Showing Results of Models using BoW

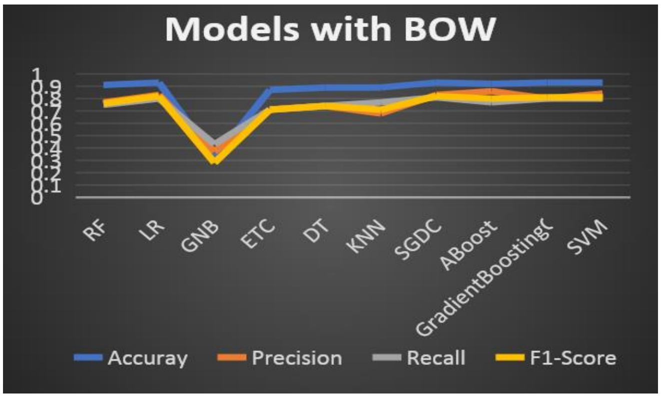

Fig. 5 Line Chart Showing the Results of Models using BoW

5. Conclusion

In this research, we harness a variety of ML techniques to address the task of categorizing user reviews. We employ feature engineering methods like TF-IDF and BoW to facilitate this classification process. A range of classifiers including: RF, LR, DT, KNN, AC, GNB, Extra Tree, SGD Classifier and SVM were trained on textual reviews. Their objective was to predict the emotional tone of user reviews, encompassing sentiments like sadness, joy, hate, anger, greed, and fear exclusively. Our findings underscore that the conclusions drawn from our tests hinge on a singular dataset that has not been previously employed for classification purposes. Additionally, the outcomes might be constrained to the specific dataset employed here. Among the models assessed, the top-performing quartet consists of SVM, LR, SGD, and GBC. These models exhibit superior accuracy, precision, recall, and F1-score metrics.

6. Future Work

Subsequent endeavors will encompass the utilization of deep learning models for conducting tests across diverse textual and categorical datasets aimed at categorizing user reviews.

References

[1] D. Garcia, C. J. Tessone, P. Mavrodiev & N. Perony, "*The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy",* Journal of the Royal Society Interface, vol. 11, 2014.

[2] J. Abraham, D. Higdon, J. Nelson and J. Ibarra, "*Cryptocurrency price prediction using tweet volumes and sentiment analysis",* SMU Data Science Review, vol. 1, no.3, 2018.

[3] D. Kahneman and A. Tversky, "*Prospect theory: An analysis of decision under risk*", *Handbook of the fundamentals of financial decision making: Part I.,* World Scientific. pp. 99-127, 2013.

[4] F.F. Bocca and L.H.A. Rodrigues, "*The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling",* Computers and electronics in agriculture, no. 128, pp. 67-76, 2016.

[5] A. Jafar, M.A. Islam, and A. Khan, "*The Effect of Social Networks and Google on Consumers' Buying Behavior in Dhaka City, Bangladesh",* Global Journal of Management and Business Research, no. 17, 2017.

[6] M. Keijsers, C. Bartneck, and H.S. Kazmi, "*Cloud-based sentiment analysis for interactive agents*", in *Proceedings of the 7th International Conference on Human-Agent Interaction,* 2019.

[7] T.H. Nguyen, K. Shirai, and J. Velcin, "*Sentiment analysis on social media for stock movement prediction",* Expert Systems with Applications, vol. 42, pp. 9603-9611, 2015.

[8] H. Sul, A.R. Dennis, and L.I. Yuan. "*Trading on twitter: The financial information content of emotion in social media*", *47th Hawaii International Conference on System Sciences*, IEEE, 2014.

[9] A. Mittal and A. Goel, "*Stock prediction using twitter sentiment analysis",* Standford University, CS229, vol. 15, pp. 2352, 2012.

[10] J. Bollen, H. Mao and X. Zeng, "*Twitter mood predicts the stock market",* Journal of computational science, vol. 2, no. 1, pp. 1-8, 2011.

[11] S.B. Kotsiantis, I. Zaharakis, and P. Pintelas, "*Supervised machine learning: A review of classification techniques.* Emerging artificial intelligence applications in computer engineering", vol. 160, no. 1, pp. 3-24, 2007.

[12] C. Lamon, E. Nielsen, and E. Redondo, "*Cryptocurrency price prediction using news and social media sentiment",* SMU Data Sci. Rev, vol. 1, no. 3 pp. 1-22, 2017.

[13] Z.H. Munim, M.H. Shakil, and I. Alon, "*Next-day bitcoin price forecast",* Journal of Risk and Financial Management, vol. 12, pp. 103, 2019.

[14] B. M. Lucey and M. Dowling, "*The role of feelings in investor decision-making",* Journal of economic surveys, vol. 19, no. 2, pp.211-237, 2005.

[15] M.M. Seif, E.M. Ramzy, and G. Abdel, "*Stock market real time recommender model using apache spark framework*", *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018),* Springer, 2018.

[16] N. Aslam, F. Rustam, E. Lee, P.B. Washington and I. Ashraf, "*Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model",* IEEE Access, vol. 10, pp. 39313-39324, 2022.

[17] J. Heaton, "*An empirical analysis of feature engineering for predictive modeling*", in *SoutheastCon 2016,* IEEE, 2016.

[18] S. Robertson, "*Understanding inverse document frequency: on theoretical arguments for IDF",* Journal of documentation, vol. 60, no. 5, pp. 503-520, 2004.

[19] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G.S. Choi, "*Tweets classification on the base of sentiments for US airline companies",* Entropy, vol. 21, no. 11, pp. 1078, 2019.

[20] D. Shah, H. Isah, and F. Zulkernine, "*Predicting the effects of news sentiments on the stock market*", *IEEE International Conference on Big Data (Big Data),* IEEE, 2018.

[21] B. Spilak, "*Deep neural networks for cryptocurrencies price prediction*", Humboldt-Universität zu Berlin, 2018.

[22] E. Stenqvist and J. Lönnö, "*Predicting Bitcoin price fluctuation with Twitter sentiment analysis",* 2017.

[23] P.C. Tetlock, "*Giving content to investor sentiment: The role of media in the stock market",* The Journal of Finance, vol. 62, no. 3, pp. 1139-1168, 2007.