Corroboration of Twitter Sentiment Analysis and Event Analysis of Indian Budget 2022 on Bitcoin Market

Abhinand G (abhinandganesh2001@gmail.com)
Anna University
Uma Maheswari V
Guru Nanak College

Research Article

Keywords: Sentiment Analysis, Event Study, Bitcoin, Union Budget, Machine Learning

Posted Date: April 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1515523/v1

Corroboration of Twitter Sentiment Analysis and Event Analysis of Indian Budget 2022 on Bitcoin Market

Abhinand G[1*†] and Dr V Uma Maheswari[2†]

[1]Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, 600025, Tamil Nadu, India.
[2]Associate Professor and Head - Department of MBA, School of Management, Guru Nanak College, Chennai, 600042, Tamil Nadu, India, https://orcid.org/0000-0002-0620-9628

*Corresponding author(s). E-mail(s): abhinandganesh2001@gmail.com;
Contributing authors: umaagaanesh@gmail.com;
†These authors contributed equally to this work.

Abstract

This study aims to understand the corroboration of the results of Sentimental analysis with Event Analysis on the Indian Budget 2022 announcement on cryptocurrency. Sentimental analysis tries classify the tweets on Bitcoin on the budget day into Positive or Negative and deciphers the overall sentiment of the bitcoin investors on the Budget Day and the subsequent 2 days. This has been made possible by using a supervised machine learning algorithm. Event Analysis identifies whether any abnormal returns are seen on the event day and the subsequent 10 days. This study observes that the range between the positive and negative sentiments is minimal and there is no abnormal returns found post the budget announcement in the event study. It may be concluded that positive sentiments nullified the negative sentiments where the overall sentiments have a mean subjectivity score that is more leaning towards the less-opinionated side, which may be accounted for the lack of abnormality was found on the event day as well as the adjustment period.

Keywords: Sentiment Analysis, Event Study, Bitcoin, Union Budget, Machine Learning

1 Introduction

Bitcoin has the highest market capitalisation in the cryptocurrency market Bitcoin is widely considered as the method of payment across countries and also by blue chip companies [8]. Bitcoin is the most active and the oldest in cryptocurrencies and other cryptocurrencies are viewed as the financial assets and not as money like bitcoin [32]. Cryptocurrency in India is unregulated and 7.3% of the Indian population own cryptocurrency [1] and the investors are expected to grow even faster [2]. The latest development in the Indian cryptocurrency market is the announcement of a 30% fixed tax rate on all income generated through crypto trading by the Indian Government during the Union Budget 2022 on 1st February 2022. "The budget

[1]https://triple-a.io/crypto-ownership/
[2]https://www.livemint.com/news/india/indians-are-spending-millions-daily-on-cryptocurrency-trading-11606906191740.html

provides clarity on taxation and shows the government's intent to take a business-friendly approach while protecting the interest of consumers and the exchequer. We hope to work with the government to help bring crypto-asset taxation at par with other asset classes and participate in the central government's vision to promote economic growth," tweeted Mr. Ashish Singhal of Coin Switch Kuber [3]. This paper attempts to analyse the sentiments of tweets by users that contain the keywords "Indian Budget 2022" and "Bitcoin" in order to establish a corroboration between the polarity of the sentiments and the presence of abnormal returns in the Bitcoin market (BTC-INR). To perform the sentiment analysis, a tweet dataset has been prepared using a scraper and after further pre-processing and application of machine learning models pre-trained by us, the polarities and subjectivities of the tweets have been observed. The paper also aims to study the effect of the range between the sentiments on the BTC-INR market with the price movements in BTC-INR through the constant return model of event study. Moreover, establishing a relationship between these two entities could give us an insight into the effect of similar events on Bitcoin markets during future scenarios with similar implications.

2 Research Questions

This study aims to answer the following research questions-

RQ1: Is there any statistically significant abnormality in the BTC-INR market during the time of the Indian Budget 2022 announcement?

RQ2: How do the positive and negative sentiments pertaining to the Indian Budget 2022 and Bitcoin affect the BTC-INR market?

The below objectives were set in order to answer the research questions-

- To conduct an event study on BTC-INR to estimate the abnormal returns using Constant Mean Return Model.
- To classify the sentiments of the tweets on Bitcoin pertaining to Indian Budget 2022 using pre-trained models
- To compare the results of sentiment analysis with the results of event analysis.

We present our paper in the following order - Section 3 deals with review of literature pertaining to event study and sentimental analysis. Sections 4 and 5 discusses in detail the event study data set and the analysis. From Sections 6 to 10, we give a detailed analysis on sentimental analysis on the bitcoin tweets on cryptocurrency related announcements in the Indian Budget 2022. Section 11 deals with discussion of the study and Section 12 deliberates on the scope for future research.

3 Related Work

Makrehchi, M et. Al [20] has attempted a model estimating sentiment based on twitter posts and use the sentiment to predict future stock market movement. Sentiments of twitter posts have a significant relationship with stock returns and volatility [29]. Ranco G, et. Al [25] found that sentiment polarity in twitter posts impact the cumulative abnormal return. Cryptocurrencies liquidity increases or decreases after positive and negative news announcements [34]. The event study on news of cryptocurrency thefts on cryptocurrency prices reveal that price rises when there is a theft news [5]. Bitcoin Investors behave rationally for positive events and a bit more irrationally for negative events [8] . In the short-term, the strength of a trend and, hence, the price in both bullish and bearish markets is invariant to volume changes; however, the volume is sensitive to price changes, especially for the upward trend. The detected relationship is unidirectional, meaning that positive information from the cryptocurrency market encourages investors to make a stronger entrance into the market, which causes bubbles and further drives price increases. Investors buy more when there is positive information from the cryptocurrency market which drives the prices to go up [32]. Tweets of Elon Musk on cryptocurrency has a significant positive impact on abnormal returns and trading volume of dogecoin and not for bitcoin [2].

Apporv et. Al [1] made use of three types of models including the unigram model, feature based model and a tree kernel model in order to classify tweets into positive, negative and neutral based on its sentiments. They identified that the sentiment analysis process used for twitter

[3] https://economictimes.indiatimes.com/markets/cryptocurrekency/articleshow/89278494.cms

data is equal to that used for other genres. They also identified that a feature analysis is essential for the revelation that important features involve the combination of polarity of words along with their speech tags. L. Lin et. Al [17] analysed the sentiments of retweeting comments and depicted the turning point in sentiment when a tweet gets retweeted. They used two primary approaches for classification which included the Support Vector Machine (SVM) and Lexicon-based method. The latter gave a higher precision and recall, which led to the conclusion that the said method is more effective in classification of sentiments.

Parabowo et. Al [24] studied the application of sentiment analysis towards unstructured data like movie reviews and comments on the social media application-MySpace. They made use of multiple approaches involving the Natural Language Processing (NLP) approach, Unsupervised approach and the Machine Learning approach. To use the machine learning approach, they primarily made use of the Support Vector Machine (SVM) classifier. They also proposed a novel approach in which each classification model could contribute to other models to have an increased level of effectiveness. Medhat, W. et. Al [21] conducted a literature survey and categorised the research papers according to the techniques used by them to conduct the sentiment analysis. They also discussed the related fields relating to sentiment analysis like transfer learning, building resources and emotion detection. Hussein, D. [13] conducted a literature survey on the challenges faced by individuals conducting sentiment analysis. They also discussed the relationship between the review structure and challenges faced in the sentiment analysis process.

[12] studied and presented comparisons of eight sentiment analysis techniques. They also studied multiple methods of sentiment analysis techniques including supervised machine learning algorithms and lexical approaches. They developed a novel method that is a combination of existing methodologies to provide the best coverage results. They observed that even with the decrease in accuracy and precision with the surge in combination of methodologies, the evaluation metrics remain in a reasonable range which is an F-score greater than 0.7. This was an indication that combination of every single methodology isn't the best way to go to achieve high accuracies and the right combination of methods varies with the kind of data that is being dealt with. Fersini, E et. Al [11] discussed the limitation that is associated with sentiment classification, where text is considered as a unique source of information. They proposed a novel method of Approval Network in order to enable the representation of the contagion on social networks in a better manner. Through experiments conducted, they concluded that the sentimental analysis methodologies based upon the proposed Approval Networks highly outperform the traditional approaches to sentiment analysis.

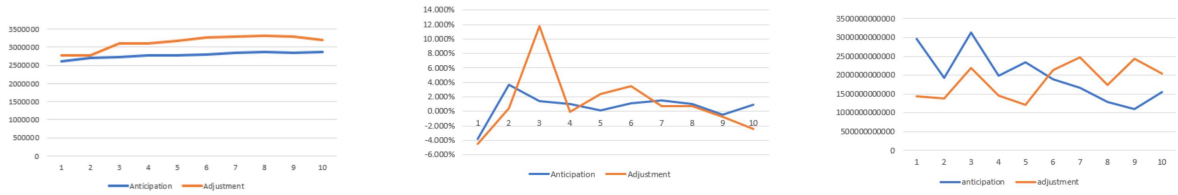
4 Event Study Methodology and Dataset

Table 1: Information on Data Set for Event Analysis

Event Date - Indian Budget '22	01-02-22
Anticipation Days: 22-01-22 to 31-01-22	10 days
Adjustment Days: 02-02-22 to 11-02-22	10 days
Estimation Window: 24-09-21 to 21-01-22	120 days
Average Return of the Estimation Window	-0.076%

This study is aimed at finding out the impact of Budget Day announcement on Crypto Currency in terms of abnormal returns by using the constant mean return model. The constant mean returns model often yields results as that of sophisticated models [6] .T-test is used to identify the presence of abnormal returns. The study focuses on BTC-INR historical data downloaded from yahoo finance [4]. A 120-day estimation window has been utilised as a test period to estimate the variance [10]. The anticipation and adjustment period of 10 days have been taken before and after the event day as the days prior and after the event day in order to capture the price effects of the announcements [19]. The average return of the estimation window was -0.076% (Refer Table 1).

[4]https://finance.yahoo.com/quote/BTC-INR/history?p=BTC-INR



(a) Closing Prices (b) Constant Returns (c) Volume

Fig. 1: Variation between Anticipation and Adjustment

5 Analysis of Data

Indian Budget 20222 was considered as the event day. 120 days prior to the budget day plus 10 days of anticipation period was taken as the estimation window. Expected Return and Standard Deviation were calculated using the bitcoin prices in the estimation window for the constant mean return model. Abnormal Returns were measured for the constant mean model using the formula -

$$AR_{i,t} = R_{i,t} - \bar{R}_i \tag{1}$$

The cumulative average residual method (CAR) measures the abnormal performance as the sum of each month's average abnormal performance [16]. The CAR starting at time t1 through time t2 where horizon length L = t2 - t1 + 1 is

$$CAR(t1, t2) = \sum_{t=t1}^{t2} AR_t \tag{2}$$

T-test was carried out for the Cumulative Abnormal Return to assess whether any abnormality in returns were detected due to the budget announcement on cryptocurrency (Refer Table 2).

The hypothesis of the event study is-

H0: *The Indian Budget announcement on cryptocurrency did not have any impact on the BTC-INR market*

Based on the T-test results of CAR for Event Day, Anticipation period and Absorption period, all the T-values are below 1.96 indicating that there is no evidence of abnormal returns due to the Budget announcement on cryptocurrency. Abnormal returns (Positive or Negative) would have been evinced in the adjustment period if the bitcoin investors have reacted in a largely polarised manner to the budget announcements. Hence, the null hypothesis is accepted.

Though the closing prices of the adjustment period (10 days post 01st February 2022) are slightly higher than the anticipation period (10 days prior to 01st February 2022), the constant mean returns for the above stated period are fluctuating. The bitcoin constant returns on 2nd and 3rd of February were slightly lower and there was a momentary moderate spike on 4th of February followed by similar lower returns in both anticipation and adjustment period. With respect to volume of trading, the anticipation period has a larger volume of trading than adjustment period until 6th of February 2022 post which the volume gained a momentum in the adjustment period - which was until the 6th of February 2022 (Refer Fig. 1).

Table 2: Analysis of Data

Standard Deviation	SD[1]	3.2%
	SD[1] (10 days)	10%
	SD[1] (21 Days)	15%
Return	Event	1.009%
	Anticipation	6.533%
	Adjustment	11.860%
	Total	19.401%
T-test	Event	0.3116295
	Anticipation	0.64
	Adjustment	1.1588065
	Total	1.308
p-value	Event	0.7558674
	Anticipation	0.5244899
	Adjustment	0.2488552
	Total	0.1933469

[1]Standard Deviation

6 Sentiment Analysis

Sentiment analysis is a technique common in Natural Language Processing (NLP). It is primarily focused on classification of text into categories like "positive", "negative" or "neutral" [1]. This paper dwells on the study of sentiments with

respect to tweets with keywords including "Budget 2022" and "Bitcoin". We have made use of machine learning based classification for sentiment analysis. Machine learning based classification is effective and practical owing to their ability to achieve a level of accuracy that can commensurate to human experts [28]. To collect a diverse set of tweet data to conduct the sentiment analysis, three different datasets are obtained from two consecutive days from the date of budget announcement, which happens to be 1st February 2022. The sentiments of the tweets from each of these days are analyzed and compared with the extent of abnormality observed in the Bitcoin market. Due to the fact that tweets are highly unstructured, there is a necessity for preprocessing the data before it is used for analysis and learning. Figure 2 explores the steps involved in the whole sentiment analysis procedure used in this research.

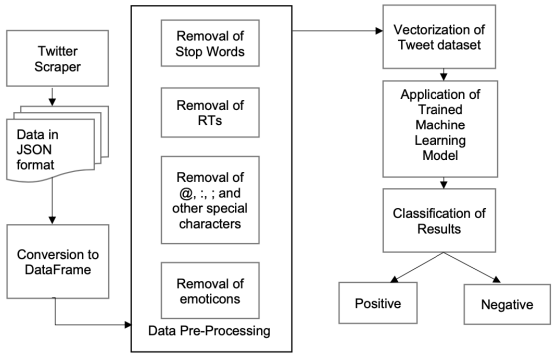


Fig. 2: Architecture of Sentiment Analysis

As observed from Figure 1, pre-processing of the dataset by us has involved a variety of methodologies to ensure better accuracy of the machine learning models. Firstly in order to train the models, pre-classified tweets have been taken from the Sentiment140 dataset [14], where tweets have been labelled as 0 for having a negative sentiment and 4 for having a positive sentiment. After the dataset is imported into a python environment, various pre-processing techniques are used. Algorithm 1 summarizes the whole data cleaning process, which involves various methods right from removal of stop words, URLs, "RT" from tweets and also special characters. These characters and strings don't play an important role in the NLP process for sentiment analysis and are even referred to as "unnecessary noise" [23]. Thus, removing them leads to a better accuracy in predicting the sentiment of a tweet.

Algorithm 1 Cleaning a tweets dataset

```
function CLEAN_DATASET
    return " ".join([word for word in str(text).split() if word not in stop_word_list)]
end function
function REMOVE_URLS
    return re.sub('(https?://[^s]+)',' ', df)
end function
function REMOVE_RETWEETS
    return re.sub('RT @[w]+',' ',df)
end function
function REMOVE_SPECIAL_CHARACTERS
    return re.sub('[^A-Za-z]',' ',df)
end function
```

7 Feature Extraction

The feature extraction technique used for this dataset is the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF is a very popular approach and has use-cases in multiple fields including information retrieval and text-mining. They are used in evaluation of the relationship of each word in a collection of multiple documents [15]. The TF-IDF value varies with respect to the frequency of a word in a document. It is based on two statistical methodologies-namely Term Frequency and Inverse Document Frequency. The term frequency refers to the frequency of the term in a document with respect to the total number of words in it.

$$Tf(word_i) = \frac{No\ of\ occurences\ of\ word_i}{Total\ number\ of\ words} \quad (3)$$

When the Tf value is high, the word is meant to have a high importance in the documents [15]. Further, inverse document frequency is a references to the rarity or frequency of a word throughout the documents. If the IDF score is high, it is an indication that the word is rarely occurring in the documents.

$$iDf = log\frac{Total\ Number\ of\ documents}{Count\ of\ documents\ with\ word_i} \quad (4)$$

Finally in order to calculate the TF-IDF value, the Tf and iDf scores are multiplied.

$$tfidf = Tf * iDf \quad (5)$$

8 Usage of Machine Learning Models

8.1 Support Vector Machine

The support vector machine was initially proposed by Vapnik et Al [4] as a supervised learning algorithm and was later again introduced by Cristianini N et Al [7]. It is used for various purposes involving regression, classification etc. Before application of the SVM model, the data needs to be vectorized and the linear SVM model object is made use of. The main objective of the SVM algorithm is the detection of a hyperplane that distinctly classifies data-points. Once the hyperplane has been identified, the data-points on either side of it are classified. Table 3 shows the type of hyperplane possible with respect to the number of input features provided.

Table 3: Types of Hyperplanes

Number of input features	Hyperplane
2	Line
3	2-D Plane

In order to ensure the least possible error during classification of two types, the optimal separating surface is made use of. It is also denoted as the largest class interval. In Figure 3, the line l refers to the hyperplane, support vectors a1, a2 and a3 refer to the positive sentiments and support vectors b1 and b2 refer to the negative sentiments.

8.2 Bernoulli Naive Bayes

The Naive Bayes Classifier makes use of the Bayes theorem. It is based upon the primary assumption regarding the independence of a feature in comparison to other features in a class. This is more

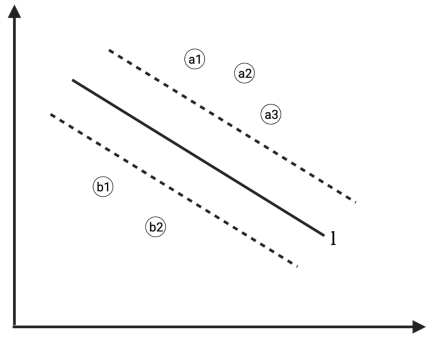


Fig. 3: Graphical Representation of SVM

preferable in some cases compared to other supervised learning algorithms owing to its simplicity and ability to quickly train a dataset. This ultimately leads to a lower time of computation [33]. The bayes theorem allows for the calculation of posterior probabilities. [22]

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{\sum_{B'=1}^{C} P(A \mid B')P(B')} \quad (6)$$

In the above equation, A refers to the sentiment labels that have been associated with each tweet, B refers to the class of sentiments that have been used. In our case, there are two classes - positive and negative. P(A | B) is the Bayesian probability of when an instance A occurs in a particular class for each value of B.

Although the Bayes classifier is widely used, it is generally sub-optimal for non-linearly separable concepts and can only learn linear discriminant functions, as observed in many researches [27] [9].

8.3 Logistic Regression

This is a model which makes use of multiple dimensions to predict and give a result for new input that has been given to it [31]. A number of independent variables are fed into the model along with their dependent counterparts to train the model. The dependent variable can either be 0 or 1, depending on the sentiment. The model also helps us analyze the effect of the independent variables on the dependent variable. The model is analytically depicted by the following equation [30]-

Table 4: Evaluation metrics for Positive and Negative Tweets

Model	Negative Tweets			Positive Tweets		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
SVM	0.81	0.79	0.80	0.80	0.82	0.81
BNB	0.81	0.78	0.79	0.81	0.78	0.79
LR	0.83	0.80	0.81	0.80	0.83	0.82

$$f(z) = \frac{1}{1 + e^{-z}}, z \epsilon R \tag{7}$$

As tempting as it is to include multiple input independent variables into the logistic regression model, it could lead to high standard errors [26]. Moreover, if the input variables are highly correlated, the preciseness of the logistic regression model greatly reduces [26]. These limitations must be taken into account while using the model.

9 Evaluation

To compare the different models, various evaluation metrics have been used. Accuracy is defined as the ratio between the number of correct predictions made by the model to the total number of predictions made by the model. Table 5 compares the accuracies of each model.

Table 5: Comparison of accuracies of various models

Model	Accuracy
Support Vector Machine	80.51%
Bernoulli NaÃŕve Bayes	79.6%
Logistic Regression	81.4%

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

In the above equation, TP(True Positive) refers to when the class of the tweet's sentiment is positive and the model predicts it to be positive, TN(True Negative) refers to when the class of the tweet's sentiment is negative and the model predicts it to be negative, FP(False Positive) refers to when the class of the tweet's sentiment is negative and the model predicts it to be positive and FN(False Negative) refers to when the class of the tweet's sentiment is positive and the model predicts it to be negative.

With respect to the accuracies of the trained and tested machine learning models, the order of best models with respect to their accuracies is as follows: Logistic Regression > Support Vector Machine > Bernoulli Naive Bayes. Precision is an estimation of how many of the positively predicted samples are actually correct and Recall is an estimation of the proportion of positive samples which were correctly identified.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

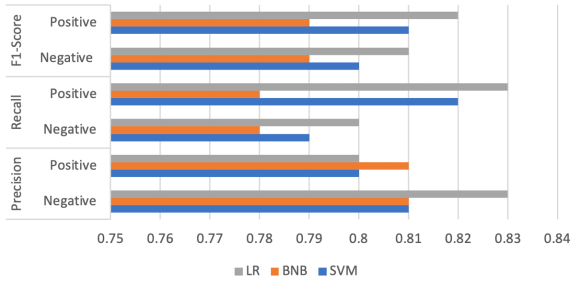


Fig. 4: Comparison of the supervised learning classifier models with respect to their precision, recall and F1-scores

Once the Precision and Recall have been understood, the F1-score can be calculated. It is mathematically defined as the harmonic mean of Precision and Recall.

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (11)$$

Thus, Logistic Regression is used for predicting sentiments of the main dataset needed for this paper owing to its highest accuracy. The evaluation metrics are for the models predicting negative and positive sentiments are given in Table 4 and Figure 4 represents them graphically.

10 Analysis of the Sentiments

In order to observe the trend in positive and negative sentiments, the dataset has been split into three categories with respect to the date of tweeting. The Logistic Regression model has been used separately in each of these categories and the results have been observed. It has been almost a common trend on all three days where not much variation between the positive and negative sentiments are observed.

The polarity isn't high enough for one sentiment to over-power the other. On February 1st 2022 -which happens to be the day of the Budget-2022 announcement, the negative tweets seem to have an edge over the positive tweets by occupying 64% of the total tweets. However, on February 2nd 2022, this gap between the positive and negative tweets appear to become narrower with the positive and negative tweets occupying 42.3% and 57.6% respectively-indicating that the overall sentiment among the twitter folk is equivocal in nature.

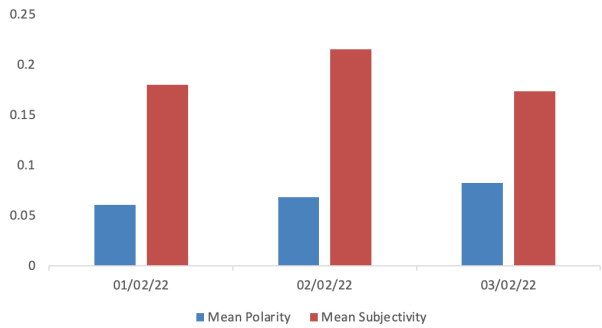


Fig. 5: Comparison of Mean Polarity and Subjectivity on the basis of the date

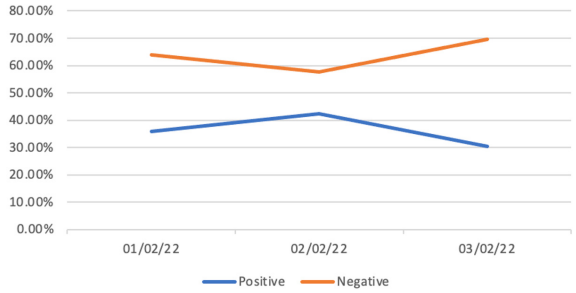
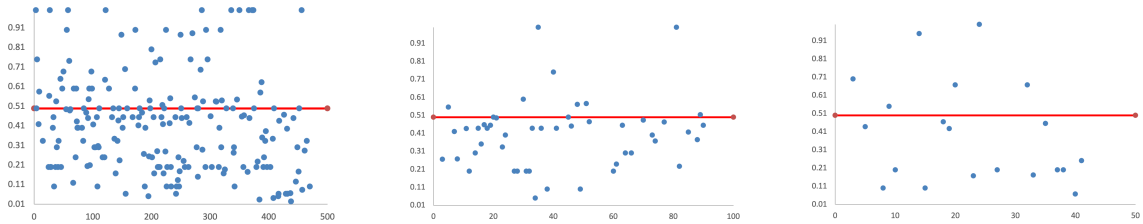


Fig. 6: Trends in Tweet Sentiments

The total number of tweets collected on each day are 470, 92 and 46 respectively-indicating that the people's interest and opinion towards the topic is highly declining. It has been observed that the average polarity of the tweets on February 1 2022, February 2 2022 and February 3 2022 are 0.0607, 0.0684 and 0.0827 respectively - where each of these values are overwhelmingly close to zero - indicating a more neutral sentiment rather than a strong positive or a strong negative sentiment. Figure 5 graphically depicts the average polarity and subjectivity scores observed on each day. Figure 6 represents the trend in the percentage share between positive and negative tweets. Further, the lack of a wide range between the positive and negative sentiments can further be elucidated by the mean polarity values of the tweets on each day.

10.1 Subjectivity

Subjectivity Classification deals with classifying a document as to how opinionated it is [18]. It is thus, a reasonable understanding that the more subjective a document is, the larger the influence of emotions, opinions and personal views in it. The analysis conducted by us in this part of the research deals with the sentiment analysis of tweets that are related to "Indian Budget 2022" and "Bitcoin". Knowing the subjectivity of each tweet is essential as it would give us a valuable insight into the kind of tweets that are being analyzed and how influential they could be towards the BTC-INR market. The value for subjectivity lies in the range [0,1] where 0 denotes the least subjective and 1 denotes the maximum possible subjectivity. When the subjectivity score is approaching 0, it can be understood that the



(a) February 1, 2022 (b) February 2, 2022 (c) February 3, 2022

Fig. 7: Distribution of Subjectivity of Tweets

document is more leaning towards the factual side [3]. For our research, we have made use of the TextBlob toolkit to calculate the subjectivity scores of the tweet dataset on each observed day. It was observed that the mean subjectivity scores on February 1 2022, February 2 2022 and February 3 2022 were 0.17965, 0.21521 and 0.17329 respectively. Figure 7 depicts the distribution of subjectivity scores on each day using a scatter plot, and a line at y=0.5 has been taken to be the reference line, where any subjectivity below the line are considered to be towards the factual side and subjectivity above the line are considered to be towards the opinionated side.

An interesting observation from Figure 7 is that on all three days more, than half of the subjectivity scores lie below the reference line, indicating that more than half of the dataset of tweets can be considered to be leaning towards a factual structure rather than being opinionated [3]. Considering the fact that the gap between the positive and negative tweets isn't very significant and the mean subjectivity of tweets on all three days approach a less opinionated side, it is a reasonable conclusion that this may be a plausible explanation for the lack of abnormality in the BTC-INR market during the recorded periods, as observed in this paper.

11 Discussion

As per the event study on bitcoin price movements (BTC-INR) during the day of the Indian budget 2022 as well as 10 days prior and post 1st February 2022, it is found out that no abnormal returns were evinced on the said period. It is similar to the result of Ante L[2] on their study on the impact of Elon Musk's twitter activity on cryptocurrency, where the results revealed that price effects were not significant for Bitcoin. In our study, we could find the negative sentiments expressed in tweets on BTC-INR were slightly higher than the positive tweets on all three days and not substantial (Refer Fig. 6) and the difference went even more narrow on 2nd February 2022 revealing that the positive sentiments were negated by the negative sentiments. Moreover, the lack of strong opinion of both the positive and negative tweets can further be explicated by distribution of subjectivity of tweets. Here, the lack of subjectivity in the observed tweets are concluded due to the plots, where more than half of the subjectivity scores fall below the target value y=0.5 (Refer Fig. 7). When we compare the results of event study and sentiment analysis, we can corroborate that sentiments expressed in twitter forums affect the financial market. This result is in tandem with the results of different studies by [29] [25] [34]. In our case we may conclude that the twitter sentiments did not result in any abnormal returns in the BTC-INR market as the positive and negative sentiments negate one another, as similarly shown in [2] and we have further expanded the cause by observing the lack of subjectivity in both the positive and negative tweets.

12 Scope for Future Research

Our research is robust as we have adopted a comparative analysis of multiple machine learning models in classifying the sentiments of twitter posts and have chosen Logistic Regression which has the highest accuracy - while at the same time comparing the models with evaluation metrics that are even beyond accuracy scores, where Logistic Regression still holds to be the best model. We have studied the linkage between sentiment analysis and event study in order to know

whether sentiments expressed get translated in the market movement. We have restricted only to BTC-INR as bitcoin is the most traded cryptocurrency in India [5]. Future studies may be extended to other cryptocurrencies and their movement in the market in the Indian. With respect to event studies we have gone for an estimation window of 120 days and 10 days prior and post the event day. The analysis can be carried out in a minute by minute basis to understand the short term movements in the price. The Sentiment analysis can also be done in three parts - Anticipation period, Event day and Adjustment period to decipher any abnormality in a detailed manner.

Declarations

- **Conflict of interest/Competing interests:** The authors declare that they have no conflict of interest.
- **Funding:** The authors declare that no funding or support from any organization has been received for pursuing this research.
- **Author contributions:** Abhinand G and Uma Maheswari V wrote the main manuscript text. V Uma Maheswari took care of the Event Study and Abhinand G was involved in the Sentiment Analysis of Twitter data, and the other sections were contributed by both the authors equally and collectively. Both the authors reviewed the manuscript thoroughly.
- **Research involving Human Participants and/or Animals:** The authors declare that there has been no such research involving human participants and/or animals.
- **Informed consent:** The research did not involve any human participants.
- **Availability of data and materials:** The datasets used for this analysis has been obtained from https://www.kaggle.com/datasets/kazanova/sentiment140 and https://finance.yahoo.com/quote/BTC-INR/history?p=BTC-INR

References

[1] Agarwal A, Xie B, Vovsha I, et al (2011) Sentiment analysis of twitter data

[2] Ante L (2021) How elon musk's twitter activity moves cryptocurrency markets. SSRN Electronic Journal https://doi.org/10.2139/ssrn.3778844

[3] Bhagat KK, Mishra S, Dixit A, et al (2021) Public opinions about online learning during covid-19: A sentiment analysis approach. Sustainability 13(6). https://doi.org/10.3390/su13063346, URL https://www.mdpi.com/2071-1050/13/6/3346

[4] Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. ACM Press, pp 144–152

[5] Brown MS, Douglass B (2020) An event study of the effects of cryptocurrency thefts on cryptocurrency prices. 2020 Spring Simulation Conference (SpringSim) pp 1–12

[6] Brown SJ, Warner JB (1985) Using daily stock returns: The case of event studies. Journal of Financial Economics 14(1):3–31. https://doi.org/https://doi.org/10.1016/0304-405X(85)90042-X, URL https://www.sciencedirect.com/science/article/pii/0304405X8590042X

[7] Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods

[8] Diaconasu DE, Mehdian S, Stoica O (2022) An analysis of investors' behavior in bitcoin market. PLOS ONE 17(3):1–18. https://doi.org/10.1371/journal.pone.0264522, URL https://doi.org/10.1371/journal.pone.0264522

[9] Duda RO, Hart PE (1973) Pattern classification and scene analysis / Richard O. Duda, Peter E. Hart. Wiley New York, URL http://www.loc.gov/catdir/enhancements/fy0607/72007008-t.html

[10] Dyckman T, Philbrick D, Stephan J (1984) A comparison of event study methodologies using daily stock returns: A simulation approach. Journal of Accounting Research

[5] https://www.forbes.com/advisor/in/investing/top-10-cryptocurrencies-in-india/

22:1–30. URL http://www.jstor.org/stable/2490855

[11] Fersini E, Pozzi FA, Messina E (2016) Approval network: A novel approach for sentiment analysis in social networks. World Wide Web 20(4):831âĂŞ854. https://doi.org/10.1007/s11280-016-0419-8

[12] Goncalves P, Araujo M, Benevenuto F, et al (2013) Comparing and combining sentiment analysis methods. pp 27–38, https://doi.org/10.1145/2512938.2512951

[13] Hussein DMEDM (2018) A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences 30(4):330–338. https://doi.org/https://doi.org/10.1016/j.jksues.2016.04.002, URL https://www.sciencedirect.com/science/article/pii/S1018363916300071

[14] Kazanova (2017) Sentiment140 dataset with 1.6 million tweets. URL https://www.kaggle.com/datasets/kazanova/sentiment140

[15] Kim SW, Gil JM (2019) Research paper classification systems based on tf-idf and lda schemes. Human-centric Computing and Information Sciences 9(1). https://doi.org/10.1186/s13673-019-0192-7

[16] Kothari SP, Warner JB (2007) Econometrics of event studies

[17] Lin L, Li J, Zhang R, et al (2014) Opinion mining and sentiment analysis in social networks: A retweeting structure-aware approach pp 890–895. https://doi.org/10.1109/UCC.2014.145

[18] Liu B (2010) Sentiment analysis and subjectivity, pp 627–666

[19] MacKinlay AC (1997) Event studies in economics and finance. Journal of Economic Literature 35(1):13–39. URL http://www.jstor.org/stable/2729691

[20] Makrehchi M, Shah S, Liao W (2013) Stock prediction using event-based sentiment analysis. pp 337–342, https://doi.org/10.1109/WI-IAT.2013.48

[21] Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4):1093–1113. https://doi.org/https://doi.org/10.1016/j.asej.2014.04.011, URL https://www.sciencedirect.com/science/article/pii/S2090447914000550

[22] Murphy KP, et al (2006) Naive bayes classifiers. University of British Columbia 18(60):1–8

[23] Patel R, Passi K (2020) Sentiment analysis on twitter data of world cup soccer tournament using machine learning. IoT 1(2):218–239. https://doi.org/10.3390/iot1020014, URL https://www.mdpi.com/2624-831X/1/2/14

[24] Prabowo R, Thelwall M (2009) Sentiment analysis: A combined approach. Journal of Informetrics 3(2):143–157. URL https://EconPapers.repec.org/RePEc:eee:infome:v:3:y:2009:i:2:p:143-157

[25] Ranco G, Aleksovski D, Caldarelli G, et al (2015) The effects of twitter sentiment on stock price returns. PLOS ONE 10(9):1–21. https://doi.org/10.1371/journal.pone.0138441, URL https://doi.org/10.1371/journal.pone.0138441

[26] Ranganathan P, Pramesh C, Aggarwal R (2017) Common pitfalls in statistical analysis: Logistic regression. Perspectives in Clinical Research 8:148–151. https://doi.org/10.4103/picr.PICR_87_17

[27] Rish I (2001) An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM New York, pp 41–46

[28] Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. https://doi.org/10.1145/505282.505283, URL https://doi.org/10.1145/505282.505283

[29] Souza TTP, Kolchyna O, Treleaven PC, et al (2015) Twitter sentiment analysis applied to finance: A case study in the retail industry. ArXiv abs/1507.00784

[30] Stanisz A (2007) Przystepny kurs statystyki zastosowaniem statistica pl tom 3

[31] Strzelecka A, Kurdys-Kujawska A, Zawadzka D (2020) Application of logistic regression models to assess household financial decisions regarding debt. Procedia Computer Science 176:3418–3427. https://doi.org/https://doi.org/10.1016/j.procs.2020.09.055, URL https://www.sciencedirect.com/science/article/pii/S1877050920319505, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020

[32] Szetela B, Mentel G, Bilan Y, et al (2021) The relationship between trend and volume on the bitcoin market. Eurasian Economic Review 11(1):25–42. https://doi.org/10.1007/s40822-021-00166-5

[33] Yang T, Qian K, Lo DCT, et al (2015) Spam filtering using association rules and naÃŕve bayes classifier. In: 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), pp 638–642, https://doi.org/10.1109/PIC.2015.7489926

[34] Yue W, Zhang S, Zhang Q (2021) Asymmetric News Effects on Cryptocurrency Liquidity: an Event Study Perspective. Finance Research Letters 41(C). https://doi.org/10.1016/j.frl.2020.101799, URL https://ideas.repec.org/a/eee/finlet/v41y2021ics1544612320316135.html