# Abstract

Bitcoin is one of the most trending crypto currency due to its price fluctuation phenomena. On the other hand, scientist now recognizing the power of predicting the sentiments base on the tweets for different events, political crisis and specially for economy. The sentiment of tweets on Tweeter about crypto currency directly or indirectly recognize the overall behavior of Bitcoin. In the proposed work, we train different machine learning models a deep learning model (BERT) for the sentiment prediction of tweets about Bitcoin. The model takes the tweets and predict their sentiment to analyze future behavior of Bitcoin. We used the Random Forest, K Nearest Neighbor and Support Vector machine algorithms for the classification of tweets. We also used the pre-trained BERT model from the transformer library for sentiment classification that has the highest accuracy.

Keywords: Bit Coin, Classification, BERT, Sentiment Analysis, Machine Learning, Natural Language Processing (NLP)

# Introduction

A cryptocurrency is type of the digital property intended to be used as medium of the exchange. The Cryptography is used to safeguard bitcoins transactions. Mostly cryptocurrencies have become one of most often utilized money for the transactions between INTERNET users for a variety of the purposes owing to its secured nature. Despite the fact that cryptocurrency values were not high in market [1], individuals utilize it for the transactions since it is secured. The Bitcoins cryptocurrency, which was first established in the 2008, was based on digital transactions in between the senders and receivers which are disseminated via a peer-to-peer system [2]. Bitcoins has become one of most valued currencies, with market capitalization is over 150$ billion [3]. Computers programmer, businessmen, Liberals, and crooks are among four groups of Bitcoins users [4]. Peoples from all over this world explain the specifics of Bitcoins currency on INTERNET because of its relevance and widespread use. Participants on a social-media sites, for example, discuss the benefits and drawbacks of the Bitcoins currency, as well as their preferences and disliking for a cryptocurrency specifically Bitcoin. These debates demonstrate the significance of a Bitcoin as well as simple popular perceptions of a Bitcoin and either someone likes or detest it. Twitters users, along with other digital media platforms, frequently discussing Bitcoin. Monitoring of people's attitudes toward Bitcoin from various perspectives can assist in determining people’s behavior is about Bitcoin and other cryptocurrencies. Understanding people’s behavior is about Bitcoin as well as an importance of the Bitcoin in respect of a people's views can be accomplished by analyzing a people's thoughts toward Bitcoin from various perspectives.

It is also assists in bringing awareness into the people’s opinions towards Bitcoin, which may be utilized to identify weaknesses and securities issues within Bitcoin networks. Furthermore, sentimental analysis towards Bitcoin tweeting can assist in comprehending reasonable problems towards the money, which can be utilized to enhance the Bitcoins structure, consequently analyzing users’ sentiments towards Bitcoins is quite important. Web Contents Mining Web Structures Mining, as well as Web Usages Mining are the three basic methods of a social web analytics. The contents analysis is generated by a social online user is referred to as web contents mining. The mainly differences between the traditional webs (also called as World Wide Web or a Web 1.0) moreover also Social Webs (also called as a Web 2.0) is that in a Web 1.0, webs owners are contents generators while ordinary peoples are contents consumers just, although in a Web 2.0, ordinary peoples are both contents consumers as well as contents generators. Instagram, Twitters, LinkedIn, Flickr, Vimeo, and other digital social media platforms are among the most popular. Web structures mining is concerned with the analysis of web-sites that are related to one another. The social structures also address how a social web related users are connected to one another. Social online usage analytics examines how a social webs users interact with social media outlets. The huge importance of a social-web has prompted many researchers to focus on it, resulting in emergence of the number of latest research topics. Opinions mining [5], also called as sentimental analysis, is one of most important research-fields in this area.

Subjectivity analysis [6], sentimental polarity and sentimental valences analysis [7], and hybrid viewpoint categorization [8] are subcategories about opinion-mining. Sentimental analysis is also commonly employed in another social web-related study domains, including such identifying prominent bloggers [9]. The identifications of important users as well as bloggers has been an prominent research domain on a social-web [10] due to importance of finding top peoples who can assist everyone else in making good life based decisions in various domains for example politics, societal problems, electronic marketing, and online commerce [11]. These problems as well as applications highlight an importance of the opinion-mining in realm of the data-mining.

Additionally, since the amount of the data available on internet grows at the exponential rate [12], text data processing for sentimental analysis and some other complicated natural languages processing applications becomes increasingly relevant. These jobs entail complicated natural languages processing activities, such as data mining and production using various methodologies [13]. This aids in comprehension of natural languages and the subsequent processing of the data. The Machine-learning and analytic methods are commonly used to perform data processing jobs.

While machine-learning algorithms performed better in various natural languages processing problem [14], they each get their own set of advantages and disadvantages.

This is a hot area for scientists right now is analyzing individuals’ opinions on various issues. Considering the social or digital media-based platforms are very useful resources for analyzing a public perspective on many subjects [15], dialogues between various social or digital media-based platforms are employed for assessment. Twitters is on the other hand, is a most popular venues for individuals from all sorts of backgrounds to debate a wide range of issues, making this one of the very useful sites for sentimental analysis [16]. Investigators have also employed Twitter to locate notable personalities [17]. Sentimental analysis encompasses a number of the Natural Languages Processing (NLP) activities, which aid in a comprehension of languages for various informational retrieval applications [12]. In this scenario, sentimental analysis is utilized to learns about people’s thoughts towards Bitcoin by analyzing their Twitters conversations regarding the currency.

The sentimental classification was performed using the real-world Twitters dataset. As a whole tweet featuring Bitcoins talks, retweets as well as tweets wherein the person mentions other people, the tweets those comprise links to the outside resources, and the tweets involving details related to cryptocurrencies are completely included in dataset. These parts aid in properly understanding people's opinions even though they can reveal the way individuals react towards Bitcoin from multiple viewpoints and at various periods. Investigation of feelings.

The comments are divided into the positive as well as negative categories using a Nave-Bayes classifications technique [18]. Whereas the Nave-Bayes is the simple classification-based model which performs classifications based on the probability, it has been utilized for sentimental classification by several studies [19]. Following the extraction of sentimental analysis data, various performances evaluation measures are employed to assess the systems performance.

# Background

## Bitcoins and cryptocurrency

The cryptocurrencies are digital money created by the computer algorithm. They are unaffiliated with a part of government and have no connection to a part of national or state bank [20]. On numerous exchange sites, digital currencies can be purchased using conventional currencies, or some cryptocurrencies can be traded for the others. September 2021, there was more approximately 12500 separate cryptocurrencies was present, as according to the Coin-Market-Cap [21].

Bitcoins, which was first released in the 2008 [22], is very most famous cryptocurrency [23] but has grown in popularity over a last six years. In the previous six years, a number of the active bitcoins addresses has expanded by almost more than threefold [24]. A number of the active address for this cryptocurrency is a number of people who are actively involved in token-transfer. The Ethereum, is a second very famous cryptocurrency, has seen the similar growth. Not only has number of active bitcoins and Ether addresses expanded, but number of the unique-addresses has indeed continue enhanced. When acquiring cryptocurrencies, each address is distinctive and can merely use at first when. According to the survey published in [25], the 53% of consumers polled in February and March 2019 bought the first initial cryptocurrency just during the 2017 and 2019. According to the Google-Trends [26], searches popularity in keyword "cryptocurrencies" began to rise in the 2017 and reached in the January 2018. It has reached its furthermore level increase in the May 2021.

## Bitcoins as the currency

Tesla-company began accepting bitcoins at an end of the March 2021, five months afterwards purchasing the 1.5 billion dollars in bitcoins [27]. Nevertheless, May 12, company advertised that it will no longer take bitcoins because bitcoins mining operations are not powered by renewable based energy-sources. They believe that after about 50% of bitcoins miners are powered by renewable energies, they will resume taking bitcoins. Bitcoins miners are similar to the modern-day banks services in that they are required to complete their transactions. Microsoft, the large multinational technology corporation, has accepted bitcoins since 2014 [28] and began taking bitcoins for its Microsoft-Store in the January 2021 [29]. By end of year 2021, AMC Theater-chains will take bitcoins. The El Salvador approved the bitcoin bill on the June 8, 2021, indicating that bitcoins will be acknowledged as currency in this country. The law will take effect in the September 2021[30].

## Bitcoins and data traffic on the Internet

According to the findings in [31], there is the link amongst online interest as well as bitcoins price. Furthermore, the strong Granger simple relationship in between bitcoin marketing volume and "online attentions" continues to occur. Google searches volume index is used to measure the internet-attention. The frequency of searchers quires for the phrase is normalized against a total number of the searches for all these terms to get all search volumes index. If either various time-series have the Granger simple relationship, one of them can be used to forecast other.

## Individual Tweet Effect on the Cryptocurrencies

Elon Musks tweets are recognized to have an influence on cryptocurrency markets. The Bitcoins price surged from the $32,000 and over $38,000 after he changed its Twitter profile to include the "#bitcoin" and referred to it by a tweet: "In perspective, it was unavoidable." Elon Musk, likewise, drew a lot of attentions to a cryptocurrency Dogecoins. The ordinary trade volume per hour for the Dogecoin momentarily increased from the $1,942 approximately to $299,330 after a tweet "I exclusively buy Doge [32]."

## Numbers of Tweets

The authors examine an impact of the three elements on a bitcoin's pricing. Tweeting volume, Tweet sentiments, and Google based Trends are three. They have discovered which Google's searches volume ranking is closely associated with a price of the bitcoin, both increasing and declining. The mood of tweets did not match price increases and was not subsequently investigated in [33]. Whenever it relates to the bitcoin's realized volatility (RV) and trade volume [34], a number of tweets connected to bitcoins in previous days is the key variable. A Realized volatility is a measurement of the variation in outcomes for the investment products by assessing its rates of the returns within the specific time," according to [35]. In [36], trade volume is defined as follows: The numbers of currencies which have raised funds within a given period of time is known as trade volume.

## Sentiments on Twitter

The findings in [37] show how Twitter pleasure mood is a powerful indicator of bitcoins and other cryptocurrencies prices. The enthusiasm sentiments on the Twitter are computed by assigning the happiness scored to tweets. The pleasure of each word is assigning the value in between 1 and 9. Loving and laughing, for instance, are rated higher than the death and killing [38]. [39] acknowledged that Twitter sentiments and quantity can be used to forecast the bitcoins-price. Competing cryptocurrencies, when compared to the bitcoin, did not complete react to the tweets in same way. Prominent cryptocurrencies including the Ethereum and the ripple, for instance, did not exhibit any estimates potential when it came to Twitters opinion. The model called the VADER [40] can be used to analyze sentiments. The VADER is the sentiments analysis models based on rules. Provided the tweet, a model computes the sentimental score. The result is the score that is indifferent, either negative, or the positive. The VADER algorithm, they believe, beats individually humans’ raters. [41] analyses the ability of three various characteristics to forecast the bitcoins price. The record price, tweets volume, and mood. They argue that a sentiment expressed in tweets are the worst. This decision was drawn by using the linear-regressions to analyze correlational factors between tweets sentiments and the bitcoin price. It is crucial to be noted that all three aspects were incorporated in different systems, so combining some of them may or may not improve outcomes. [42] looked at the influence of the influential profiles contrasted to complete users. Around June 1, 2017 as well as June 25, 2018, 21 million tweets were analyzed. The little more than 17,000 total tweets were sent by 50 most powerful profiles. The influence of numerous parameters has been investigated employed the logistic system. They say which the quantity of the positive tweets, a day before has the big beneficial effect on the bitcoin's increased in prices. The same may be stated for the quantity of the negative tweets sent a day before. Surprisingly, the quantity of impartial tweets had no discernible effects on a price of the bitcoin. The similar visible effect may be seen when examining at a quantity of the positive and the negative tweets from a most prominent profile. This shows simply examining at a most prominent profiles may be adequate to see the correlation in returns, despite the fact that, this is the far smaller sample size than all bitcoin-related tweets.

# Literature survey

## Social-networks analytics

The activity of gathering data from the social networks (digital media) platforms and evaluating data to help managers, researchers, analysts and decisions-makers in addressing specific challenges is known as digital media analytics. Many individuals, particularly medical experts, engineers, and marketing business managers, have employed social networks analysts. In compared to traditional electronic media research, where data gathering and its analysis are generally performed manually, automatic social networks analytics are both, cost-effective and speedy [43]. When large digital media websites allowed businesses to find massive amounts of customers data from the sites, the attractiveness of social networks analytics skyrocketed.

One of traditional goals of social-networks analytics is to recognize and describe structure of the network, which is commonly accomplished through the use of a graph-theory [44]. Participants in the confined networks find some connections among other participants of another communities whenever data is collected about them. The present or absent of these links is significant since they determine the type of networking, including the friendship or communications system. Connection shows the types of association that exists amongst networks members. When a person recognizes alternative as a buddy, it reflects that people out-degree, as well as when change recognizes someone as a fellow, it symbolizes an individuals in-degree in social-networks analysis [45]. Every link can additionally have the weight [46], that could reflect a relationship's intensity (weaker or stronger), or some other knowledge just about relationships (kind of guidance or knowledge). "Value networks data" refers to the connections with weights [47].

Researchers gathering networks data often propose the research questions just about the broad range of areas, covering such business, sports, diseases, health-care, political, finance, advertising, and many more, instead of concentrating solely on social-networks structure. Furthermore, social-networks analytics can be used to evaluate social-networks data in order to gain fresh thoughts from consumers and strengthen their interactions.

For businesses, there are numerous open-sourced tools, professional tools, and famous platforms which provide simple as well as fundamental analytics. Scientists, inventive professionals, and decisions-makers are employing these technologies to discover latest ways of acquiring, combining, and analyzing data from social-networks platforms in order to better comprehend the business-based environment, consumers, relationships, and to develop latest products. Organizations must plan and alter their social-networks analytics efforts often in today's technology-based business climate. Nevertheless, there is shortage of categorizations which managers may use to detect different types of the analytics and access appropriate methodologies for analyzing information from the social-media (Lee, 2017).

## sentimental analysis

Sentimental analysis is the discipline which uses the natural languages processing technologies to determine individuals’ opinions and their feelings from textual data [48]. Sentimental analysis as well as textual mining, dissimilar traditional data mining technique, can deal by nonstructural data [49]. It is also referred as the opinion-mining, and it focuses on a text categorization challenge. Because of large amount of the data, collecting sentiments from the web-scale textual data can be a time-consuming and difficult operation [50]. Originally, there was little subjective data found on the websites however, with the introduction of social-media-networks in before 2000s, individuals began to share own ideas. The majority of existing sentimental analysis methods rely on categorizing individuals’ tweets either negative or positive. They are divided into two types: supervised techniques (which required training-data) and dictionaries-based techniques [51].

In the supervised approaches, such as a naive-Bayes as well as support vectors machine, concentrate on a training classifier [52]. They have produced some excellent results; nevertheless, getting the training-data, particularly for a continuously changing Twitters data, could be difficult [54]. By using distance supervised approach [53], which uses autonomously produced the training-data and uses symbols such ":-)" as well as ":(" to categorize the tweets as a positive or a negative, is one solution to overcome such challenge. Nonetheless, automated labeling of the training-data may result in the form of errors, affecting classifiers performance [55]. Additional flaw in such techniques is inherent domains dependence, which means when classifiers learned on the data from single domain (for example. Sports based tweet) generate appropriate outcome when applied to the data from another area (For example. electronic-commerce based tweets) [56].

Dictionaries based techniques, also known as the lexicon-based techniques use the set of a are coded terms to establish a semantic compatibility of the texts [57]. As according to [58], a lexicon technique is the unsupervised technique in which textual input is categorized into the collection of predetermined sentimental classes as the main sentimental analysis approach. The textual sentimental scores are computed using the sentimental lexicon, that is the dictionary of terms and associated sentimental ratings [59].

The Lexicon-base approaches are regarded frameworks to determine the polarities of textual data when it comes to data analysis. Because the terms in textual data may express the polarity that differs from that stated in a lexicon, they may not yield good efficient outcomes [60]. This difficulty can be overcome by creating context-specific lexicons to reduce terms polarity. Researchers frequently employ both of these strategies indicated previously.

## Twitter

Over time, social networking sites have grown in popularity. Some research focused on use of sentimental analysis technologies in users’ comments and articles, allowing anybody to be the prospective contents provider. In addition, the twitter is the valuable source of the information for this sort of activity. As a result, Twitter has grown in popularity between its consumers, making it a valuable resource for the analysts. Users intents on the Twitter vary; for instance, several used it solely for conversation to discussing their daily activities. Another use of it for the professional purposes. and yet others used it to propagate malicious contents. Because Twitters characters constraint of 280 characters makes it difficult for people to expressing themselves fully, some resort to using face expressions. As a result, people can convey emotions including such grief, excitement, embarrassment, rage, and some others with shorter words [61]. Twitters offers the advantage of capturing a large number of trustworthy users current attitude and behavior [62]. In compare to conventional research methodologies such as polling and comments cards, Twitter could be the latest research approach with reduced bias associated with relationships as well as recall of personal situations [63]. According to [64], collaborative understanding on the Twitter may be much more precises than different methods of retrieving information, such as opinions polls and the surveys. Nevertheless, it must be correctly examined in the high sufficient quantity. Furthermore, Twitters API provides nearly real-time accessibility to the tweets or remarks, making it the ideal environment for the large-scale nearly-real-time sentimental research [65]. Many scientists use [66] and [66] as the source for sentimental analysis.

## Crypto currency

In study [67] provides a brief overview of evolution of concept about wealth. With a rise of the merchantability and capitalistic wealth started to be comprehended as ownership of the material items such as the gold or some other important and costly metals, as well as methods to generate commerce and supply it in order to achieve more materialistic things; thus, funding was successively recognized as a source of the power. When the capitalism economy was formed in a western culture, currency has become the primary source of prosperity because of its stable earnings, which allows it to be turned into a range of abilities. The real hardcore of this world’s wealth is increasingly concentrated in the financial assets instead of the real wealth and assets. The successful billionaire’s income is often appraised largely on the market valuation of his company, rather than the values of his yacht or automobile [68]. The usage of the crypto-currencies has grown in popularity as a result of governmental and national banks failures during 2008 financial crisis [69]. The Bitcoin and some different crypto-currencies may indeed offer less expensive alternatives to conventional debits as well as credit cards systems [70], and it remains to be seen if the rise of the crypto-currencies will reshape the definition of "money." According to current research by [71], latest concepts as well as technology associated with the crypto-currencies (such as decentralization of currency and block-chain) have a potential to usher a world into the latest economy, pushing conventional economics-based transactions into a digital sphere [72].

Bitcoin was not a first and starting digital-currency; The e-Cash, e-Gold, and Flooz were all previous attempts at the purely virtual method of exchange money with the e-gold (developed by a Gold and Silver Reserves Corporation) being a most prominent. The end-to-end electronic or digital monetary systems was first stated in the brief journal research article by [73], that outlines the goal of this digital currency and how it could be produced and maintained. [64] discusses the shortcomings of current electronically payments system as well as a high costs of resolving conflicts in that mechanism. [75] says that the cryptographic demonstration would allow any two consenting participants to interaction directly with every other without a requirement for the trusted third one party to address the inherent confidence concerns in a electronically payments system. Both vendors and consumers would be protected by cryptographic evidence [76].

The goal of the digital money was to improve an existing electronically payments system by allowing people to exchanged electronic coins by using the digital signatures as a proof of the ownership. The initial Bitcoin transactions took place in the January 2009, and much more than three years, various studies predicted that the Bitcoin would be in a circulation with much more than the 6.5 million users [77]. The Bitcoin is the commodity currency combination. Bitcoin is a most well-known of payment systems. Autonomous, dispersed systems which use the shared ledger information technique called as the block-chain paired with safe cryptography to create the digital monetary or crypto-currencies which exist online [78]. As according to coin market cap [1,] Bitcoins are not only a crypto-currency; over 100 others, for example Ripple as well as Ethereum, have surfaced thus far. The cardano and lite coins in contrast to the Bitcoin, based on their market capitalization.

## Tweets Sentiments analysis about negative or positive opinion

The Bitcoins and a cryptocurrency have been one of most famous research subjects over past six years. For Bitcoins researchers focusing on the social or digital media platforms for various types of data-mining as well as knowledge-based extraction activities. Mai F looks into the role of the social-media in deciding Bitcoins values. The author claims that social or digital media perceptions are the key factor in deciding the value of Bitcoin [79]. To examine the shift in the Bitcoin prices, *Georgoula et al.* used sentimental analysis. They employed the machine-learning models to conduct sentimental analysis and analyze the cause of Bitcoins price fluctuations [80]. Matta M et al study if people's good feelings about the Bitcoins helps or hinder price of Bitcoin rising. According to the findings, there is the substantial similarity in between Bitcoins and Googles Trends about data, but they did not identify extensive sentimental analysis outcomes or the impact of the people’s opinion on Bitcoins pricing [81]. Investigators are also using sentimental analysis to forecast price of Bitcoins over the course of the day depends on public attitude [82]. In perspective of people’s sentiments, Kaminski J investigated a correlation and calamities in between Bitcoins and Twitter tweets. The paper looks into people’s sentiments as Bitcoins prices fluctuate [83]. The sentimental analysis of cryptocurrency, encompassing Bitcoins and some other cryptocurrency, was carried out by Bhargava GM et al, who looked at people’s attitudes on various cryptocurrency terminology [84]. The impacts of digital or social-media on Bitcoins are discussed by Mai F et al, who indicate which is effect of the social medias platform on Bitcoin are affecting Bitcoin on the hourly basis [85]. Furthermore, users comments and responses are useful in the determining Bitcoins and other cryptocurrency volatility [86]. The implications of users opinion on anticipating Bitcoins currency swings are discussed by Kim YB et al [87]. The author presented the currency values fluctuation algorithm based on users feelings which predicts the currency fluctuation by assessing users attitudes towards the certain currencies [88].

Let’s have a look that how important Nave-Bayes is and how extensively it has been utilized as the data-mining technique for categorization. For detecting seismic-events as well as nuclear explosions, Naive-Bayes categorization is utilized [89]. Scientists utilizing the Artificial-Immune-System developed the adapting attribute weighted for the Naive-Bayes categorization [90]. Furthermore, features weighting is combined with the Naive-Bayes classification algorithms.

In comparison to generic Naive-Bayes categorization method [91], the trials reveal that these parameters infrequently decrease the performance. The frequency related approach [92] employs the Naive-Bayes classifications techniques for the detection of DDOS assault. Ischemia stroke is also classified using by Naive-Bayes categorization and T1 normalized MRI scans [93]. Passively indoor localization-base categorization is also done by using the Naive-Bayes, with a final finding showing that an algorithm performs as well as 86 percent [94]. The negative-category based information in textual classification is also done with the naive-Bayes categorization, and results are good enough [95].

Against a substitution then attacks comparison, privacy-preserved naive-Bayes categorization techniques are used to calculate a server offline period for overall computing overhead [96].

For sentimental analysis, data analysis algorithms have been used frequently in past.

# Exploratory Analytic

## Dataset

In the proposed work, we will use the BTC tweets sentiments dataset from Data-world platform. The selected dataset is based on the tweets of different users along with their sentiments. BTC tweets sentiments dataset is generated by collecting the tweets about Bitcoin. Here we will use the BTC dataset for the prediction of tweets sentiment by using the deep learning model.

The original Dataset contains the following number of rows and Columns:

|  |  |  |
| --- | --- | --- |
| **BTC Tweets Sentiments** | **Number of Rows in Dataset:** | 50852 |
| **Number of Variables in Dataset:** | 10 |

**Dataset – Summary of Attributes**

The BTC dataset is a tabular dataset that is store in csv file format. The chosen dataset contain the total ten features/columns. The first eight features of the BTC dataset were collected manually while the next two features (New-Sentiment-Score, New-Sentiment-State) were generated with the help of NLP model. The table described the features name, description of feature, and type of data in that feature.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variable Name | Description | Type |
| 1 | ID | Serial Number / identification Number | Numeric |
| 2 | Date | Date of the tweet when tweet posted on tweeter. | Date |
| 3 | Tweet | Content of the tweet | Text |
| 4 | Screen-name | The name of the user who posted the tweet. | Text |
| 5 | Source | List of prominent words in tweet. | Text |
| 6 | Link | Http link of extracted tweet. | URL |
| 7 | Sentiment | Sentiment of the tweet in string categorical format. i.e., positive, negative | Categorical |
| 8 | Sent-score | Label encoded form of Sentiment feature. | Numeric |
| 9 | New-Sentiment-Score | Sentiment of the tweet in string categorical format. i.e., positive, negative by NLP model | Categorical |
| 10 | New-Sentiment-state | Label encoded form of New-Sentiment-Score feature. | Numeric |

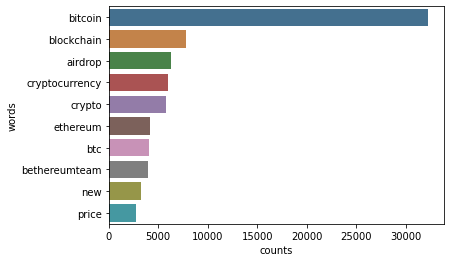
**Target Variable Description**

The BTC dataset contain the sentiment and Sent-Score that have the most suitable target variables for our proposed solution. Both features contain the sentiment of tweet in string or numeric format respectively. After the initial understanding of the dataset and by considering our problem statement, we finalize the sentiment feature as our target variable. The target variable contains the three unique values that consider as classes for the prediction of sentiment.

|  |  |
| --- | --- |
| Features | ID, Date, Tweet, Screen-name, Source, Link, |
| Target | Sentiment |
| Classes | Positive, Negative, Neutral |

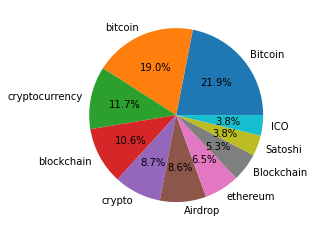
## Most Frequent Words

We perform the preprocessing steps on the tweets of dataset to found the most frequent words in dataset. For the extraction of most frequent words, we remove stop words, punctuation, hashtag sign, special characters and change them to lover case. The top 10 most frequent words is shown in Figure 1.



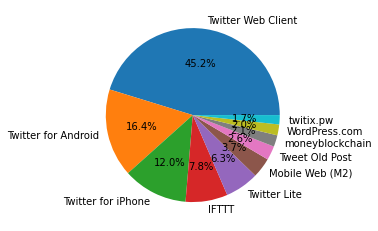
## Top 10 Hashtags

To discover the most using hash tags for the tweets of bit coin, extract all the word followed by the hash tag sign. The result of most using hash tags is shown in Figure 2.



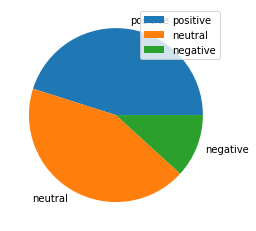
## Top Sources for Tweets

This figure shows the top 10 sources for Tweets. It seems most of them are from the Twitter Web app and Twitters from Android Systems and iPhones are somehow the same. The result of top 10 sources of tweets is available in Figure 3.



## Tweet ’s Sentiment Analysis

For the understanding of dataset, we tried to find out the number of tweets against each category. We found the three types of sentiments gains the tweets. The number of tweets for each sentiment is shown in Figure 4.



# Methodology

According to Nair et al. [2021] study, Sentiment Analysis is the Identification and extraction of writer's feelings from their material in the form of tweets by employing natural language processing (NLP) and analytic methods. The sentiment of the writer can be deduced from the words used in the text, sentiment analysis techniques, lexical dictionary of terms, and the polarity of the words. Polarity refers to whether the content is written with a positive or neutral sentiment in the writer's head. One of the most active fields in NLP is sentiment analysis, sometimes known as opinion mining. Sentiment analysis data can be used in a variety of sectors, including economics, psychology, sociology, marketing, advertising, and political science [Hutto and Gilbert, 2014].

In the proposed work, we did the sentiment analysis bit coin tweets data. We perform different machine learning and BERT algorithm. We also compare the results of all models and mentioned them in result section.

## Data Preprocessing

Preprocessing of dataset in the essential step to clean and transform data for optimal performance of machine learning models. In this regard, we use different techniques of data preprocessing like cleaning and tokenization.

For the cleaning of the data, we remove all the stop words from tweets. In addition, we also remove the punctuations, special characters, numbers, URLs from tweets dataset. Furthermore, the emojis and hash tags between the tweets text were removed and remaining text was converted into lower case. In the preprocessing phase last step was stemming that convert the different form of words into its standard form. For the stemming of tweets, we used Porter Stemmer Algorithm from NLTK tool. It also helps in the extraction of the important features. The sample of tweet after preprocessing is shown in Table 1.

|  |  |
| --- | --- |
| Tweet before Preprocessing | Tweet After Preprocessing |
| RT @ALXTOKEN: Paul Krugman, Nobel Luddite. I had to tweak the nose of this Bitcoin enemy. He says such foolish things. Here's the link: httâ€¦ |  |

1. Feature Extraction

After converting the tweets into the stem words, there is need to extract the important features for classification and remove the irrelevant words that do not have any special impact on sentiment.

According to Fouad et al. [2018], various features must be retrieved in order to represent the input tweet text. The most basic and classic technique for feature extraction is the bag-of-words (BoW). It can contain both uni-grams (single-word terms) and bi-grams (two-word terms). A bag of words is an unsorted group of words whose position is ignored but whose occurrence in the text is recorded. Lexicon-based features are another approach; for each tweet, the positive and negative terms were counted using the lexicon. The extracted words' part of speech could be saved in a pre-processing stage. PoS traits were counted in terms of nouns, verbs, adjectives, and adverbs.

TF-IDF (Term Frequency Inverse Document Frequency) and Word2Vec are two more common weighed approaches for extracting features. IDF increases the weight of words that appear rarely and reduces the weight of terms that frequently occur in a text, while TF measures how frequently a word appears in a text. The product of these scores is scored by TF-IDF. The TF-IDF assists in the identification of meaningful words that add value to text [Shofiya and Abidi, 2021].

 Based on the retrieved features, each tweet is represented as a vector of numbers. The dimension of this vector grows as the number of distinct phrases in the input tweet collection grows [Fouad et al., 2018]. For this study, we used the scikit-learn Count Vectorizer class to turn the text into a bag of words representation, which allows us to count words and their frequencies.

## Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that uses the labels of observations during  
the training or learning process. To classify or predict unlabeled words, we use these trained  
algorithms. By converting the data and drawing a hyperplane using mathematical functions  
called kernels, SVM identifies the decision boundary. There are four types of kernels: Linear,  
Non-Linear, Sigmoid, and Polynomial. The linear kernel is used for linear separation, and  
radial basis function (RBF) is used for non-linear separation [Shofiya and Abidi, 2021].

## Random Forest (RF)

Hama Aziz and Dimililer [2021] mentioned Random Forest constructs learning classifier using  
a set of decision trees from train data. Then collect the votes from different decision trees.  
Based on this collected information it can decide the final class label for test data [Kumar and Jaiswal 2020].

1. BERT  
     
   Devlin et al. [2018] described that BERT stands for Bidirectional Encoder Representation  
   from Transformers. It is an algorithm that considers both directions from right to left  
   and left to right simultaneously while training the model. BERT uses the Mask Language  
   Model (MLM). In Mask Language Model, some tokens mask randomly from the input, and  
   predicting the original vocabulary id of the mask word based on the context is the objective.  
   Most of the previous language models were left to right.  
   BERT is bidirectional, and as a result, we can get more accurate results. In comparison  
   to other sentiment analysis algorithms, BERT can extract more features. Instead of using  
   LSTM (Long Short Term Memory), BERT uses Transformers to get the context of the words  
   because it is more attention based algorithm[Nair et al., 2021].  
   8BERT: Bidirectional Encoder Representation from Transformers  
   After pre-processing the text, the first step is to tokenize the data and vectorize it to  
   have the corresponding ids of words. During tokenizing, [CLS] token adds to the beginning  
   of the sentence, and [SEP] adds to the end of each sentence [Nair et al., 2021]. The input  
   embedding is the sum of the token embedding, the segmentation embedding, and the position  
   embedding [Devlin et al., 2018]  
   **Figure 10:** BERT Input Presentation  
   Then a small probability gives to each token by adding a small MASK value to it .After  
   that perform an activation function on it and check the score or polarity value.If this value  
   is greater than the threshold set it is positive else it may be neutral or negative [Nair et al.,  
   2021].

# Results

We explore different machine learning models for the tweet sentiment analysis. After the preprocessing of the tweets, we used the preprocessed data as input and the label of the tweets as output. To convert the text into features number, we used the Count Vectorizer. It transformed each word to a vector represented by the frequency (count) of that word that occurs in the entire dataset.

We also train the BERT model for the sentiment analysis of bit coin tweets. For BERT model, we used the BERT Tokenizer that convert the tweets into vector of words. Further, we used the embedding layer in front of out BERT model that generates the feature vector of same length by word of vectors of different length. The generated feature vector is used as input by BERT model for the training.

For the conversion categorical labels, we used the Label Encoder function of Scikit-Learn library. Label Encoder transform the negative, Neutral, and Positive labels into 0, 1, and 2 respectively. For the split of dataset into training and testing set, we used the train-test-split function of scikit-learn. By using the train-test-split function, we split the data with 70% and 30% in training and testing set respectively.

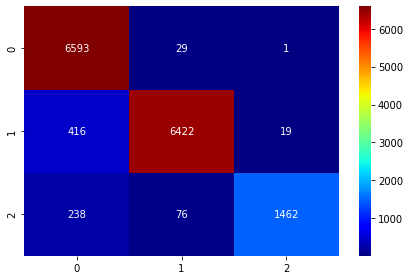
For the evaluation of the trained models, we used different evaluation measures including accuracy, precision, recall and f1-score. As Powers [2020] mentioned, the proportion of real Positive cased that correctly predict positive is called Recall or Sensitivity. Conversely, the proportion of Predicted Positive cases that are correctly Real Positives is called Precision or Confidence. So, we can mention the recall means “how many of this class you find over the whole number of elements of this class," and the precision will be “how many are correctly classified among that class." The f1-score is the harmonic mean between precision & recall. The formula if calculating the accuracy, precision, recall and f1-score is mentioned in eq 1-4 respectively.

## Random Forest

The first model that we used for the classification of bit coin tweets sentiment was Random Forest model. It is efficient to learn complex relationship between input features and output label. The classification of Random Forest model is shown in below table.

|  |
| --- |
| Classification Report for Random Forest Model |
| precision recall f1-score support  0 0.91 1.00 0.95 6623  1 0.98 0.94 0.96 6857  2 0.99 0.82 0.90 1776  accuracy 0.95 15256  macro avg 0.96 0.92 0.94 15256  weighted avg 0.95 0.95 0.95 15256 |
| Training accuracy Score : 0.9998325024925224 Validation accuracy Score : 0.9571213220925688 |

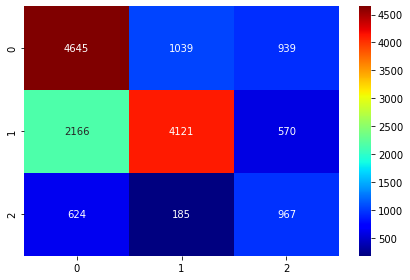
The confusion matrix of the Random Forest Model is below:



## Support Vector Machine

The second model that we used for the classification of the tweets was SVM. SVM showed the 0.64% test accuracy. The complete classification report of SVM model is show below:

|  |
| --- |
| Classification report : |
| precision recall f1-score support  0 0.62 0.70 0.66 6623  1 0.77 0.60 0.68 6857  2 0.39 0.54 0.45 1776  accuracy 0.64 15256  macro avg 0.60 0.62 0.60 15256  weighted avg 0.66 0.64 0.64 15256 |
| Training accuracy Score: 0.9998325024925224 Validation accuracy Score : 0.9571213220925688 |

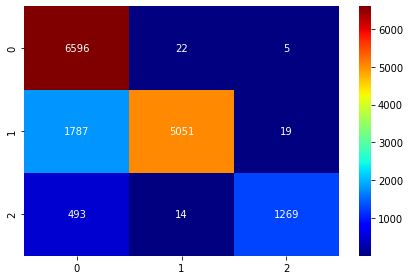
Here is the confusion matrix of support vector machine trained model:  


## KNN Model

KNN is supervised machine learning algorithm for classification and regression problem. It is simple and easy to implement algorithm for small size dataset. KNN is our third algorithm that we used for the sentiment classification of bit coin tweets. KNN showed the 0.84% test accuracy. The complete classification report of KNN model is shown below

|  |
| --- |
| Classification report: |
| precision recall f1-score support  0 0.74 1.00 0.85 6623  1 0.99 0.74 0.85 6857  2 0.98 0.71 0.83 1776  accuracy 0.85 15256  macro avg 0.91 0.82 0.84 15256  weighted avg 0.88 0.85 0.85 15256 |
| Training accuracy Score : 0.9998325024925224 Validation accuracy Score : 0.9571213220925688 |

Here the confusion matrix of KNN

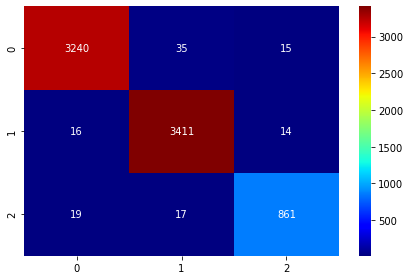


## BERT Model

Lastly, we used the pre-trained BERT model developed by the Hugging Face. Further, ThilinaRajapakse created a transformer library that used the pre-trained model by hugging Face. We trained the BERT model by transformer library for the sentiment classification of tweets. The trained BERT model showed the 0.99% training accuracy. The detailed classification report of BERT model is shown below:

|  |
| --- |
| Classification report : |
| precision recall f1-score support  0 0.99 0.98 0.99 3290  1 0.98 0.99 0.99 3441  2 0.97 0.96 0.96 897  accuracy 0.98 7628  macro avg 0.98 0.98 0.98 7628  weighted avg 0.98 0.98 0.98 7628 |
| Training accuracy Score : 0.9998325024925224 Validation accuracy Score : 0.9571213220925688 |

Here the confusion matrix of BERT model



Compare result of trained models

In or research, we got the best results by BERT model followed by the Random Forest and KNN. BERT model showed the 0.98% testing accuracy and the highest value for other evaluation measures. The comparison of all evaluation measures is also presented in the below table

-----------+----------+----------+-----------------+----------+

| Metrics | BERT | SVC | RF\_CLassifier | KNN |

|-----------+----------+----------+-----------------+----------|

| accuracy | 0.984793 | 0.637979 | 0.948938 | 0.846618 |

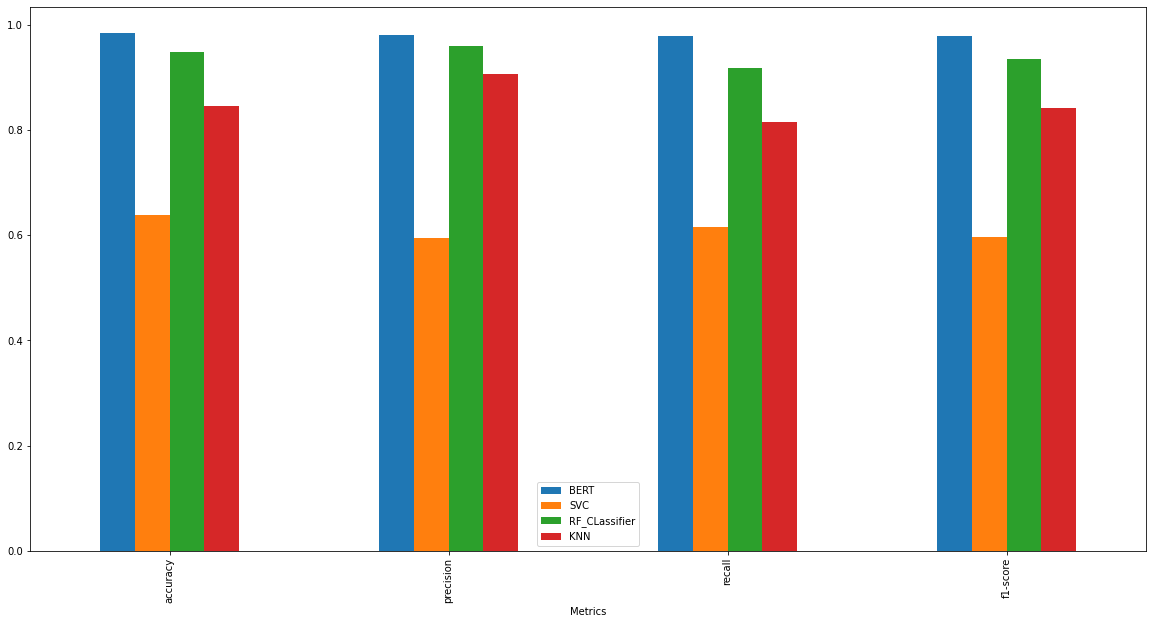
| precision | 0.980571 | 0.595433 | 0.960058 | 0.90583 |

| recall | 0.97865 | 0.615606 | 0.91841 | 0.81569 |

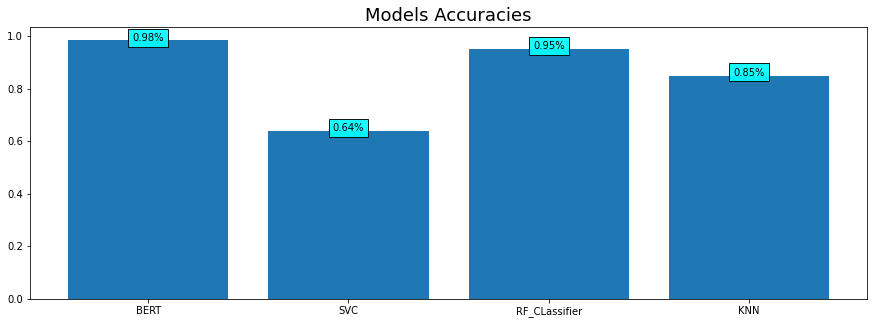
| f1-score | 0.979601 | 0.597047 | 0.93594 | 0.841304 |

+-----------+----------+----------+-----------------+----------+

The comparison of all evaluation measure is also presented by bar graph. The bar show the four groups of bars relevant to accuracy, precision, recall and f1-score for each model respectively. The bar is presented in below figure:



To elaborate the comparison of accuracy between the models and to show the best performance model, a comparative bar graph was plotted that showed the accuracies of trained model. The plotted bar graph is shown in below:



# Conclusion

In recent years, sentiment classification has been a very active research field due to social  
media and web technology advances. In this research, we tried to classify the sentiment in tweets on the topic of bit coin currency. Firstly, we select the important feature and then perform different attempts to classify the tweets against the sentiment labels of negative, neutral and positive. We tried different machine learning models and pre-trained BERT model for the classification of sentiment. We got the highest 0.98% testing accuracy by BERT model. Our trained BERT model showed the significant results over the other trained models. The prediction of bit coin price in future base on the tweets sentiment can be discovered in future studies.