Sentiment Analysis using R: An approach to correlate Bitcoin price fluctuations with change in user sentiments

Thesis submitted in partial fulfilment of the requirement for the degree of

Bachelor of Computer Science and Engineering

under the supervision of

Dr. Jia Uddin

by

Shaomi Rahman (14101181)

Jonayed Nafis Hemel (14301049)

Syed Junayed Ahmed Anta (14101105)

Hossain Al Muhee (14301070)

School of Engineering and Computer Science

April 2018

BRAC University, Dhaka, Bangladesh

Declaration

We hereby declare that this thesis is based on results obtained from our own work. Due acknowledgement has been made in the text to all other materials used. This thesis, neither in whole nor in part, has been previously submitted to any other university or institute for the award of any degree or diploma.

Signature of Supervisor

Dr. Jia Uddin

Signatures of the Authors

Shaomi Rahman (14101181)

Jonayed Nafis Hemel (14301049)

Syed Junayed Ahmed Anta (14101105)

Hossain Al Muhee (14301070)

Acknowledgement

First and foremost, we would like to thank Almighty Allah for enabling us to initiate the research, to put our best efforts and successfully conclude it.

Secondly, we would like to thank our supervisor Dr. Jia Uddin. Without his support, guidance and contribution we could not have done it. He motivated us to do this research thoroughly and was always there to help us with any aid he could offer. This paper has been completed due to his great supervision and inspiration. We are truly grateful to him.

We revere the patronage and moral support extended with love, by our parents as well as our friends. They helped us with their direct or indirect suggestions which aided in achieving our goal. We would also like to acknowledge the assistance we received from numerous resources over the Internet especially from fellow researcher's work.

Last but not the least, we thank BRAC University for providing us the opportunity of conducting this research and for giving us the chance to complete our Bachelor's Degree.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
BTC	Bitcoin
CART	Classification and Regression Tree
CC	Cross Correlation
GLM	Generalized Linear Model
ICO	Initial Coin Offering
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurring Neural Network
SL	Significance Level
SVR	Support Vector Regression
TB	Terabyte

Abstract

Analyzing sentiments has been widely regarded as a popular technique by many researchers, and Twitter nominated the most user-friendly, and reliable social media supplying the stream of sentiments. Among the trendiest topics of discussion in such social platforms, cryptocurrency, and most notably Bitcoin ranks the highest, both providing curiosity as a technology, and a lucrative asset to trade. This thesis studies the correlation among user sentiments from Twitter and the change in price of Bitcoin, to carve out a scalable model by manipulating the category of sentiments as variables and appropriate quantitative machine learning techniques.

The work finally achieved a stable precision for determining movement in price, with a high of 75% in accuracy in the short run.

CHAPTER 01

Introduction

Huge amount of data is generated every day, about 2.5 million TB of data on an average in 2017 per day [1]. Social media plays a major role in generating these data, mainly because people share every sort of feelings or emotions in social media, starting from a celebrity scandal to their success in life. The most prominent social media are basically Twitter and Facebook where 500 million tweets, and 1.8 billion pieces of information are shared every day respectively [2]. These chunks of information are useful at times, if nurtured and analyzed properly. People's post and tweet are many a times a post-expression of an event, action and sometimes determining opinions regarding finance, and politics. It is also proven by scientific research that social media has significant relation with human personality and behaviour in the actual world [3].

To gain knowledge of influences, opinions, subjectivity, assessments, sentiments, approaches, evaluation, observations, feelings, borne out in text, reviews, blogs, discussions, remarks, reactions, or some news, social media is also very essential source [4].

We intend to use sentiment regarding cryptocurrency to figure out patterns in changes of their price, and among numerous established digital currencies, Bitcoin[5] is the first introduced and has strong correlation with other digital currencies that follow. Like all other currencies, Bitcoin is also affected by socially constructed opinions; whether those opinions have basis in facts, or not [6]. Hence, we set off our research to explore these areas and figure out a definitive way to forecast price change using sentiments as variables.

1.1 Motivation

The first and foremost point of contact between two individuals is established by their exchange of sentiment, may it be a vocal, written or physical gesture, or a mere trade of the looks on the eyes. In this rapidly evolving world of technology, data to us is of no use if we can't extract meaning from them. Consequently, Sentiment Analysis has been hugely popular among

data scientists to process natural language. Accurate and precise categorization of sentiments into emotions or sentiments such as ‘anticipation’ or ‘surprise’from a sentence is far from reaching its peak, while the number of research done to categorize text into either positive, negative or neutral has comparatively exceeded the former, and hence has stirred us to try to test out an algorithm based in with R studio using the R programming language.

From our study of popular words searched over the internet, and result of word clouds formed by keywords in blogs and social media, we have concluded that the rising interests of people in recent time are on the terms ‘cryptocurrency’, ‘Bitcoin’, ‘blockchain’, ‘Ethereum’ etc. RStudio was used to find results of popular word in the form of word clouds. These results lead us to our topic of our research. However, sentiment analysis is taken a step further if properly implemented with Machine Learning algorithms to extract implicit insights of the collected data. Provided our topic of interest and its popularity, such insights are not only important to discern characteristics and dependencies, but also extremely valuable and profitable, if found scalable. The drive to figure out a precise sentiment analysis algorithm and use machine learning, where appropriate, to acquire valuable business and research insights is the primary drive of our research.

1.2 Problem statement

1. Can we categorize a twitter sentiment regarding cryptocurrency into emotions other than only positive and negative? How accurately or to what extent do they correlate to change in cryptocurrency price?
2. Can we figure out an ideal prediction model using quantitative analysis and machine learning?
3. Is the model profitable to use as business insight in the short and long run?

1.3 Thesis Outline

Chapter 1 Discusses the basic introduction, motivation and problem statement of our work.

Chapter 2 Builds on the background of methods and techniques used.

Chapter 3 Provides literary review of works related to our topic

Chapter 4 Delineates the work-flow and overall algorithm

Chapter 5 Addresses results of our work

Chapter 6 Concludes with limitations and future scopes, followed by the bibliography

CHAPTER 02

Background Analysis

2.1 Sentiment Analysis

Information in texts can be divided into two categories: **facts** & **opinions**. Facts are objective expressions and opinions are subjective expressions of people's feelings, sentiments, appraisals [9].

Sentiment analysis is a type of data mining that measures the inclination of people's opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources. The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information. It is also known as opinion mining.

Internet (especially Social Media) is the ultimate platform for people to share information and opinion. Businesses or any kind of organization can be benefited hugely by this trend of sharing opinion. Analyzing sentiment provides people's view may help a business to improve their product, it can also help political organizations to predict elections which often help candidates what needs to be done for them to win. Anything that depends on public in general can be benefited by Sentiment Analysis. Due to the tremendous value of Sentiment Analysis there are now 20-30 companies that offer sentiment analysis in United States alone [9].Furthermore, according to the TIME Magazine, Cambridge Analytica, which worked with both Trump and Texas Sen. Ted Cruz's presidential campaigns, got access to information on Facebook users and their friends through a researcher who misled the social network giant, saying the information he was gathering would be used strictly for academic research, according to recent reports and these information are considered as "Extremely valuable".

2.2 Types of Sentiment Analysis

With the increasing demand of sentiment Analysis there is a number of ways in which sentiment analysis is implemented. According to Indurkhya, Damerau these approaches can be categorized in the following categories [9]:

2.2.1 Document level sentiment classification

It analyses a text and gives the result in binary such as weather the text implements a positive or negative sentiment. The approach of this classification is: given an opinionated document *d* that comments on an object *o*, determine the orientation *oo* of the opinion expressed on *o*, that is, discover the opinion orientation *oo* on feature f in the quintuple (*o , f , so, h, t*), where *f = o* and *h, t, o* are assumed to be known or irrelevant. This classification can be done in 2 ways:

a. Supervised Learning: The text is compared with a set of positive and negative word list to calculate positivity or negativity.

b. Unsupervised Learning: It first extracts the adverbs and adjectives and then each phrase is computed in these 3 equations:

$$PMI(term1, term2) = log_2\left(\frac{\Pr(term1 \cap term2}{\Pr(term1)\Pr(term2)}\right)$$

$$oo(phrase) = PMI(phrase, \text{"excellent"}) - PMI(phrase, "poor")$$

$$so(phrase) = log_2\left(\frac{hits(phrase\ NEAR\ \text{"excellent"})hits("poor")}{hits(phrase\ NEAR\ \text{"poor"})hits("excellent")}\right)$$

These algorithms compute the average of all phrases in the review, and classifies the review as recommended if the average *oo* is positive, not recommended otherwise.

2.2.2 Sentence level subjectivity and sentiment classification

Given a sentence *s*, two subtasks are performed: 1. Subjectivity classification: Determine whether *s* is a subjective sentence or an objective sentence, 2. Sentence-level sentiment classification: If *s* is subjective, determine whether it expresses a positive or negative opinion.

2.2.3 Opinion lexicon generation

Each word is assigned a positivity and negativity value. It may lead to misleading most of the times.

2.2.4 Feature based sentiment analysis

This task can be divided into two subtasks:

1. Identify object features that have been commented on. For instance, in the sentence, "The picture quality of this camera is amazing," the object feature is "picture quality."
2. Determine whether the opinions on the features are positive, negative, or neutral. In the above sentence, the opinion on the feature "picture quality" is positive.

2.2.5 Sentiment analysis of comparative sentence

The subtasks of this task are:

1. Identify comparative sentences in *d* and classify the identified comparative sentences into different types or classes.
2. Extract comparative opinions from the identified sentences. A comparative opinion in a comparative sentence is expressed with (*O1, O2, F, PO, h, t*) where *O1* and *O2* are the object sets being compared based on their shared features F (objects in *O1* appear before objects in *O2* in the sentence) *PO* is the preferred object set of the opinion holder h t is the time when the comparative opinion is expressed.

2.2.6 Opinion spam and utility of opinion:

Sometimes some texts may be misinterpreted as opinion-based text which do not express any opinion or any subjective expressions. In this classification, the text is classified weather the text in fact shares any opinion or shares facts.

With a lot of researches, a number of approaches are available for Sentiment Analysis. Cambria, Havasi & Hussainhas categorized the methodology of Sentiment Analysis into followings [10]:

- **Keyword Spotting:** text is classified into categories based on the presence of fairly unambiguous affect words [12]
- **Lexical Affinity:** This approach assigns arbitrary words a probabilistic affinity for a particular topic or emotion [13]
- **Statistical Methods:** These methods calculate the valence of keywords, punctuation, and word co-occurrence frequencies on the base of a large training corpus [14]

2.3 Machine Learning:

Machine Learning is a branch of Artificial Intelligence[27] which is the current big thing that every major technology company is focusing on. Technology is in a state where research is happening to execute every task using machine learning. Machine Learning is a computational process where machine learns itself how to solve a particular problem. Machine Learning algorithms can be supervised or unsupervised. Some of the Machine Learning Algorithms that have been used in our project are described below.

Regression is being used in this research to statistically find the relationship between two variables. The variables here are human sentiment and price of the cryptocurrency. One of the variables is independent and the other is dependant. The independent variable is the sentiment of the people which is extracted from the posts of the popular social media ‘Twitter’. The dependant variable, on the other hand, is the price of the desired cryptocurrency. We want to

know how human sentiment affects the change in the price of cryptocurrency. Let the independent variable be denoted as 'x' and the dependent variable be 'y'.

2.3.1 Linear Regression

Linear regression attempts to model the relationship between two variables by setting a linear equation to the observed data. If a graph of y against x is drawn, a best fit line can be drawn which will be a straight line with y-intercept of b_0 having a slope of b_1.[27]

Equation: $y = b_0 + b_1 * x_1$

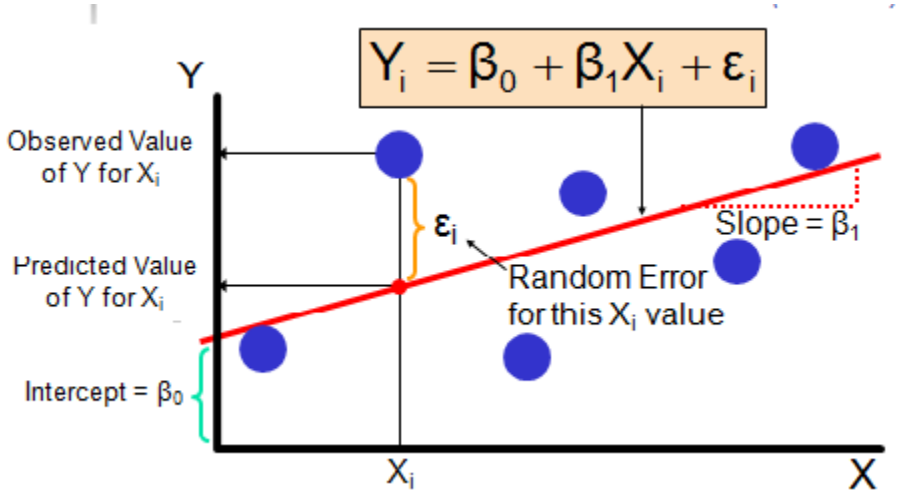

Figure2.3.1: Linear Regression Graph[27]

2.3.2 Multiple Linear Regression

Multiple linear regression is similar to linear regression but with more than two variables. There is only one dependent variable 'y' and more than one independent variables 'x_1, x_2, …, x_n'.[27]

Equation: $y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$

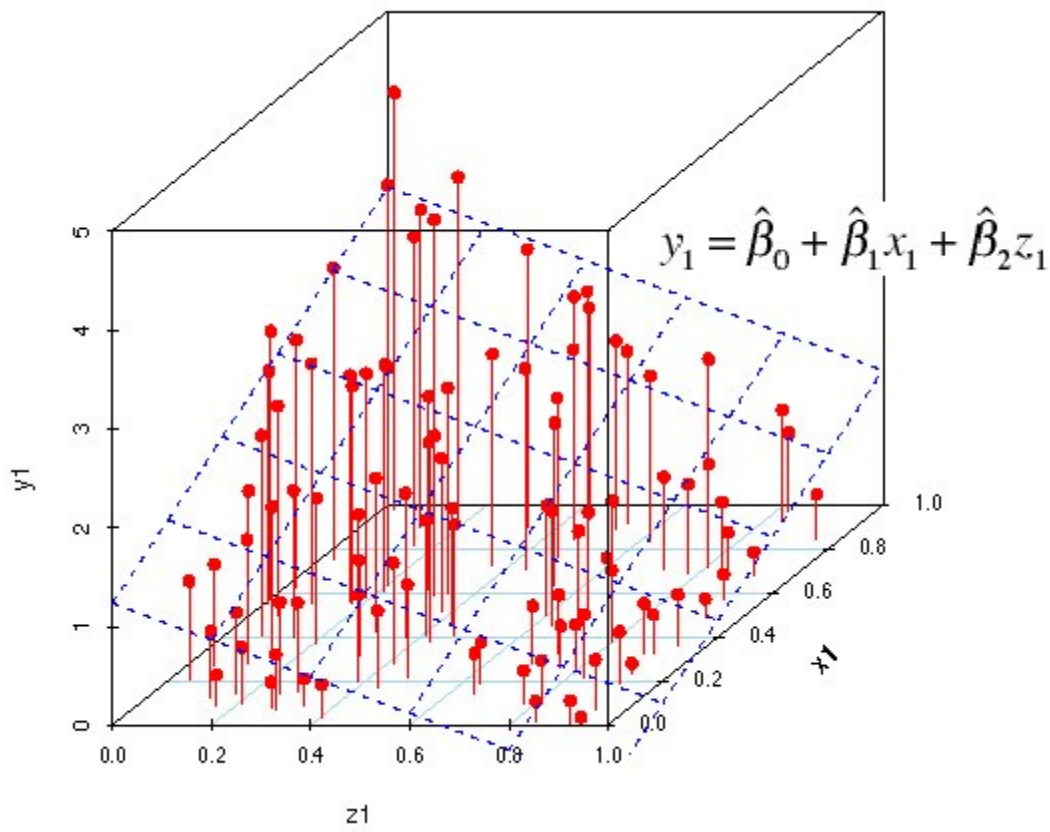

Figure 2.3.2: Multiple Linear Regression[27]

There are mainly 5 methods of building a multiple linear regression model namely all-in, backward elimination, forward selection, bidirectional elimination and score comparison. Of these methods, only backward elimination was used to build the model.

Step 1: Select significance level (SL)

Step 2: Put all the variables (predictors) into the model

Step 3: Consider the predictor with highest P-value. If P > SL, advance to Step 4, else Stop.

Step 4: Remove the variable with highest P-value.

Step 5: Rebuild the model without the removed variable. Go to Step 3.

2.3.3 Polynomial Linear Regression

Polynomial Regression is considered to be a special case of linear regression where there is no linear relation between dependent and independent variable. So, to fit the data perfectly independent variables are manipulated with different power. In this model, instead of different variables, the same independent variable 'x' is used but in different powers. Here, the line of XY is not a straight line but it is a curve.[27]

Equation: $y = b_0 + b_1*x + b_2*x^2 + \ldots + b_n*x^n$

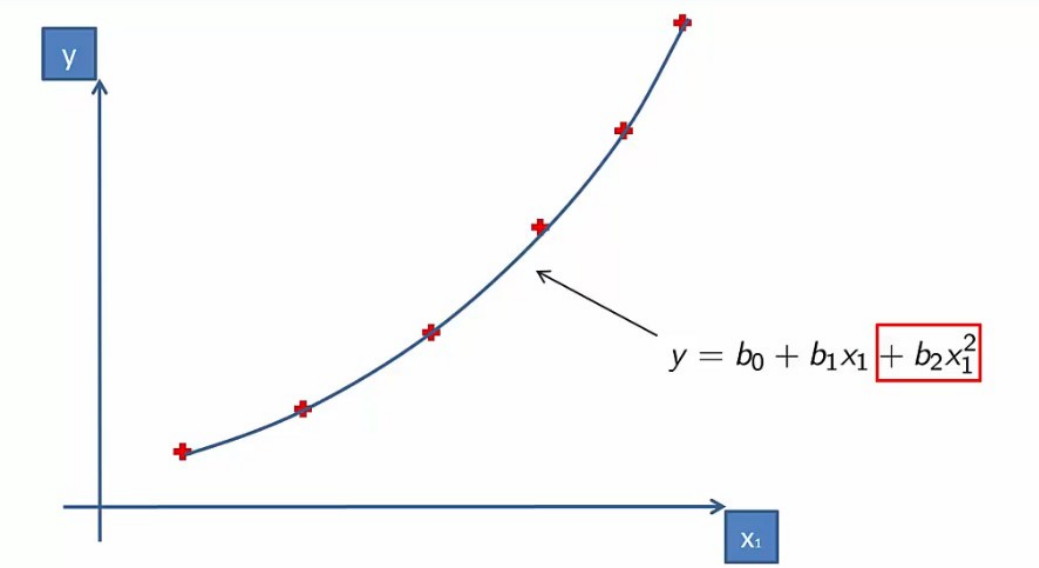

Figure 2.3.3: Polynomial Linear Regression[27]

2.3.4 Support Vector Regression (SVR)

Support Vector Regression is suited for the extreme cases. SVR draws a decision boundary known as a hyperplane near the extreme points in the dataset. SVR algorithm is used to best segregate two different classes.[27]

Build decision function of the form: $f\left(\underset{x}{\rightarrow}\right) = \underset{w}{\rightarrow}\underset{x}{\rightarrow} + b$

$$\text{Find } \underset{w}{\rightarrow} \text{ and } b \text{ that}$$

$$\text{Minimize } \sum_{i=1}^{N} \max\left(0, \left|y_1 - f\left(\underset{x}{\rightarrow}\right)\right| - \varepsilon\right) + lambda|| \underset{w}{\rightarrow} ||_2^2$$

Loss

Penalty

Figure 2.3.4 (a): Support Vector Regression Equation

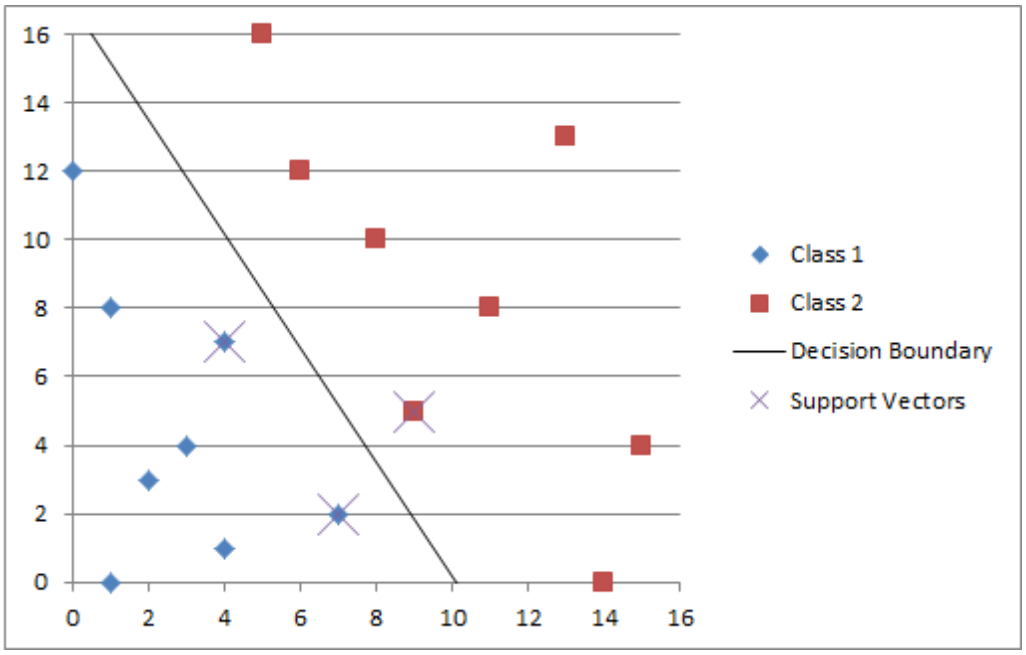

Figure 2.3.4 (b): Binary Classification of SVR model

2.3.5 Decision Tree Regression (CART)

Decision tree is a type of supervised learning algorithm which is mostly used in classification problems. A decision tree is a flowchart like structure where each internal node denotes a test on an attribute, each branch represents an outcome of a test and each leaf or terminal node holds a class label. CART can handle both numerical as well as categorical data.

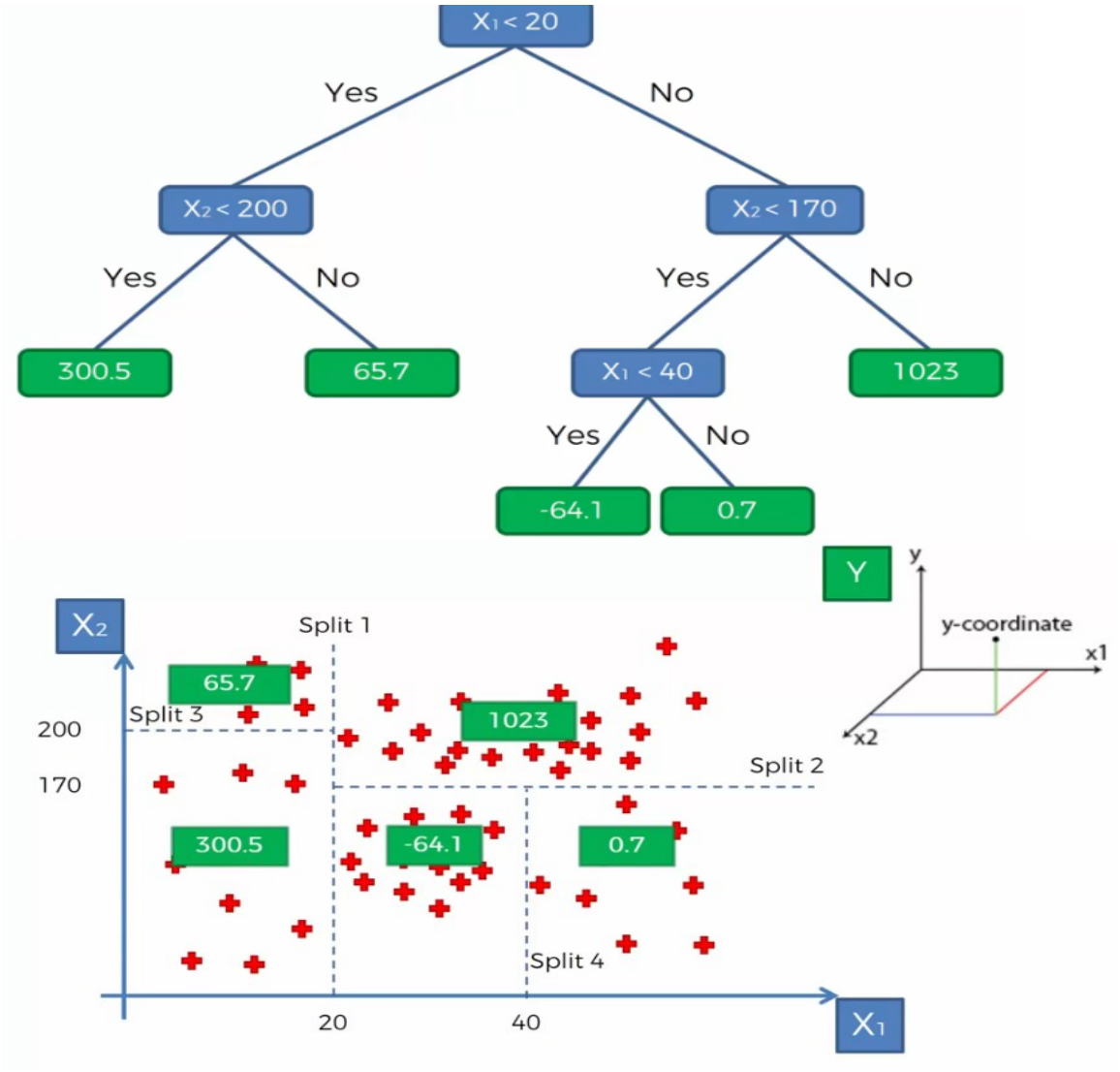

Figure 2.3.5: Decision Tree Model Representation in Graph [27]

2.3.6 Random Forest Regression

Random Forest Regression develops lots of decision trees based on random selection of data and random selection of variables. It provides the class of dependent variable based on a 'forest' of trees. The more trees there are in the forest, the more robust will be the prediction and thus higher accuracy. To classify a new object based on attributes each tree gives a classification and we receive the tree votes for that class. The forest chooses the classification having the most

votes over all the other trees in the forest and in the case of regression it takes the average of the outputs by different trees.[27]

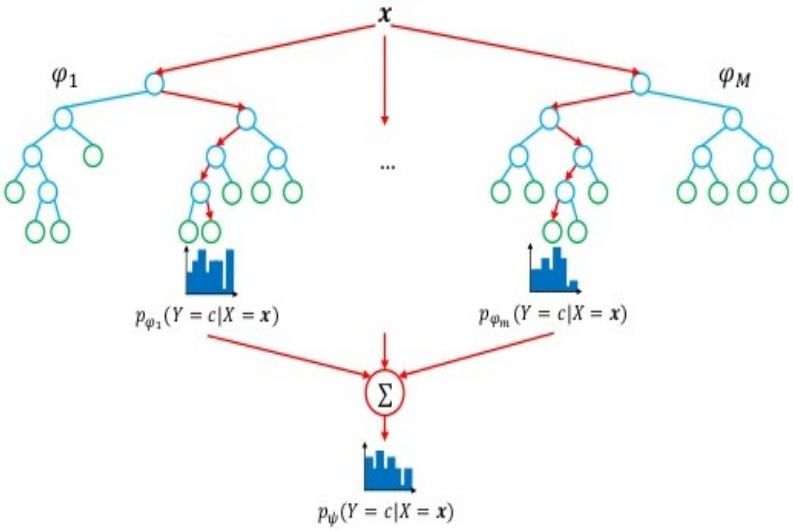

Figure 2.3.6: Random Forest Regression [27]

2.4 Literature Review

Cryptocurrency uses encryption and verification techniques to regulate the generation of units of digital currency, operating independently in a decentralized peer-to-peer network. During recent years, this topic has given rise to numerous speculations, controversies and interest in finance and politics. However, there exists ambiguity in technological definition, it is stated as a protocol, platform, currency as well as payment method[16]. Digital currencies or cryptocurrencies is trending on google trend and is among the most talked topic in the world of social media, due to its characteristics, unpredictability and lucrative rate of price, and hence calls for a deeper economic analysis.

The maiden and widely established cryptocurrency is known as Bitcoin which was created in 2009 by Satoshi Nakamoto and is termed as an decentralized, independent electronic cash-system [5] have already been subject to a lot of criticism due to the facility it provides for money-laundering, black-market transactions such as “The Silk Road”, and terrorist financing. However, there are tons of positive effects to it as well just like all the coins have two sides. Recently, in February 2017, the bitcoin price soared again to cross the $1000 psychological threshold and has continued to do so [15]. Many authors claim that as more and more people are accepting Bitcoin transaction over time, the volatility is decreasing slowly and the time is near when it will mature, hence creating the need to further study. Moreover, high volatility means there is a provision of higher profits for business insights regarding the predictability of crypto-assets.

Bitcoin transactions are secured by blockchain technology, creating blocks in an open ledger with transaction information encrypted and verified by miners who lend out their computing power in return of a small reward in bitcoin. [20] suggests that there are a myriad of data regarding these assets which have not yet been effectively utilized to provide a reflection of their correlation with the price, as we find almost 250,000 transactions taking place per day, with more than 40 exchanges with 33 different currencies accepting BTC [17]. Shah et al.[19] used latent source model as developed by Chen et al.[21] for price prediction. They claimed to reach around 87% accuracy, but when their work was tried to replicate, failed vigorously as it was reported that the author may have used hand-picked data instead of all available data to obtain

satisfying result. Geourgoula et al.[22] tried to show dependency with sentiment analysis using support vector machines and has found positive relation with Wikipedia views and network hash rate. Matta et al.[23] looked for dependencies between BTC price, tweets and Google Trend, discovered moderate to weak correlation as well, however was limited to study of only 60 days. Furthermore, there are arguments stating their tweets weren't properly pre-processed, and provision for spam or fraudulence like sentiment manipulation using fake news weren't handled[17]. Kristoufek[24] found positive correlation with search-engine views, network hash and mining difficulties with Bitcoin price in the long run using Wavelet coherence. Greaves et al.[26] analyzed blockchain data and found limited predictability of BTC price using SVM and ANN because BTC price cannot be determined by blockchain data alone. Similarly, Madan et al.[25] attempted the same using SVM, Random Forest and Binomial GLM and reported accuracy over 97%, but there is proof of lack of cross-validation[17]. So, the generalization of such a model is of doubt.

CHAPTER 03

Proposed Model

3.1 Data processing

3.1.1 Gathering real-time tweets

Twitter's API was used in combination with twitteR and RCurl to collect the required data (Tweets) for the sentiment analysis. The 'twitteR' is an open source framework written in R which facilitates tweet collection from Twitter's API. 'RCurl' is another framework for R platform which helps the data to be extracted from a website's database. Filtering based on hashtags or words are carried out by twitteR and this is considered as an efficient way of collecting relevant data. Filter keywords such as 'bitcoin', 'cryptocurrency', 'Satoshi', 'Nakamo', 'BTC' and 'XBT' were used to tighten the search further to only include Tweets related to Bitcoin.

```
oauth_endpoint(authorize = "https://api.twitter.com/oauth" , access = "https://api.twitter.com/oauth/access_token")

download.file(url = "http://curl.haxx.se/ca/cacert.pem" , destfile = "cacert.pem")

reqURL<- 'https://api.twitter.com/oauth/request_token'

accessURL<- 'https://api.twitter.com/oauth/access_token'

authURL<- 'https://api.twitter.com/oauth/authorize'


consumerKey = "**************************************************"

consumerSecret = "**********************************************"

accesstoken = "**************************************************"
```

```
accesssecret = "**************************************************"

Cred <- OAuthFactory$new(consumerKey = consumerKey ,

                         consumerSecret = consumerSecret ,

                         requestURL = reqURL ,

                         accessURL = accessURL,

                         authURL = authURL)

Cred$handshake(cainfo = system.file('CurlSSL' , 'cacert.pem' , package = 'RCurl'))

save(Cred, file = 'twitter authentication.Rdata')

load('twitter authentication.Rdata')

setup_twitter_oauth(consumer_key = consumerKey , consumer_secret = consumerSecret
, access_token = accesstoken , access_secret = accesssecret)

some_tweets = searchTwitter("cryptocurrency" ,n = 500 , since = "2018-03-20" , until="
2018-03-21" , lang = "en" )
```

Listing: Example function that gathers a stream of filtered tweets

3.1.2 Reducing the noise from dataset

Irrelevant tweets can be undesirable for the research. Some strategies were followed to correct the automatically generated tweet

- Tags, hashtags & mentions are removed with regular expression
- Links of other URLs are removed
- Also, unnecessary and multiple punctuations are removed
- All the letters were transformed into lower case

- Spelling mistakes were fixed
- Extra spaces were removed

3.1.3 Dataset Processing:

Each week we took 500 tweets and did sentiment analysis using RStudio library sentimentR and sentimentAnalysis. We compared sentiment and price change using different regression model and sentiment analysis package.

3.2 Algorithm

We have implemented a number of regression model to build a relation between sentiment and price change. For each regression model we have manipulated the dataset in different way. The processes are described below:

3.2.1 Linear Regression

In linear regression there has to be one independent value. So, using SentimentQDAP method we came up with a sentiment score which represent if the sentiment is positive or negative or neutral. And if the sentiment is positive or negative then with a decimal value the amount of positivity and negativity is computed. To compute the sentiment score of each week the average sentiment score of all 500 tweets is taken and then compared with the price change using linear regression method. Then the data is split into training and test set in ⅔ ratio. Then linear regression is applied into the training set and the hypothesis generated is measured in test set. caTools library is used to implement the algorithm.

3.2.2 Multiple Linear Regression

In this regression model there will be multiple independent variables in which dependent value will be predicted. Here each independent variable is considered as a feature. To fit data in the regression model we have used sentimentR library to determine the sentiment value of a tweet. get_nrc_sentiment method is used.This method measures 500 tweets and divides the

sentiment into 10 variables. Those variables are anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, positive. After adding each week's price and sentiment the data is put into Multiple Regression Model. In this model also caTools library is used.

3.2.3 Polynomial Regression

To prepare dataset for Polynomial Regression there had to be one independent variable. So, for that, we prepared our dataset the same way dataset for linear regression dataset was prepared. But there are two small changes. First, we did not need to divide the dataset into test set and training set. Secondly, the independent variable had to the power of 2. Then the dataset is implemented in Polynomial Regression algorithm.

3.2.4 Support Vector Regression

SVR will be implemented using decimal sentiment score and price change and there will be no test and training set. Library e1071 is used to implement the algorithm.

3.2.5 Decision Tree Regression

The method of preparing dataset for SVR and Decision Tree Regression will be the same. In this algorithm library rpart is used.

3.2.6 Random Forest Regression

The dataset preparation is same as the previous one. Then the regression algorithm is implemented using randomForest library in R.

3.3 Flowchart

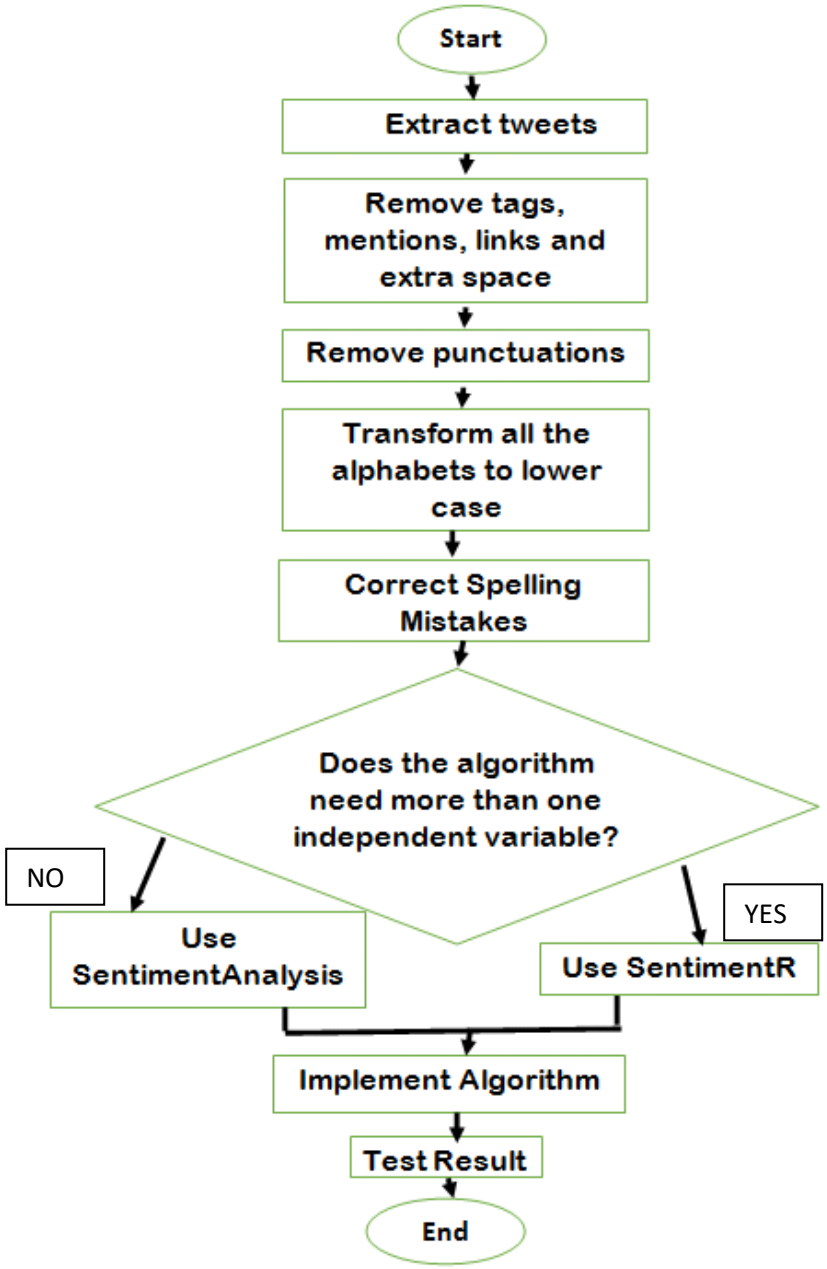

Figure 4.3: Flowchart of the Whole Algorithm

3.4 Tools Used

The whole program for the research was written in R using RStudio along with Microsoft Excel to store and process the data set.

CHAPTER 04

Experimental Results

After implementation and computations our findings are the followings:

4.1 Linear Regression

The training data fits the hypothesis in 10% accuracy. and for test set some of the results are:

TABLE I: Linear Regression Results

Actual Result	0.92	0.53	-0.74	-6.09	8.74	-0.02	3.48	0.00
Predicted Result	-0.25	-0.51	-1.29	1.31	-3.1231	-3.906	-1.81	7.84

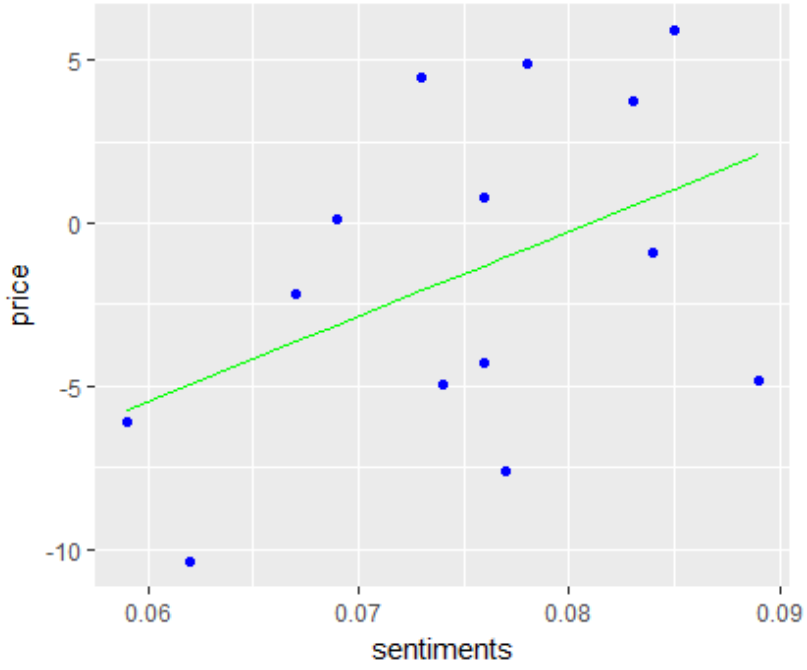

Figure 5.1:Result of Linear Regression Testing[27]

4.2 Multiple Linear Regression

The training set fits the hypothesis 42.4%. By backward elimination process that variables that have high significance are anger, anticipation,fear, joy, sadness, positive. The results of the test set are:

TABLE II: Multiple Linear Regression Results

Actual Result	0.92	0.53	-0.74	-6.09	8.74	-0.02	3.48	0.00
Predicted Result	-25.096	3.757	10.61	-4.24	16.619	-11.99	4.26	-6.230

4.3 Polynomial Linear Regression

We figured that taking the power of x upto 5 fits the dataset mostly which is 7%.

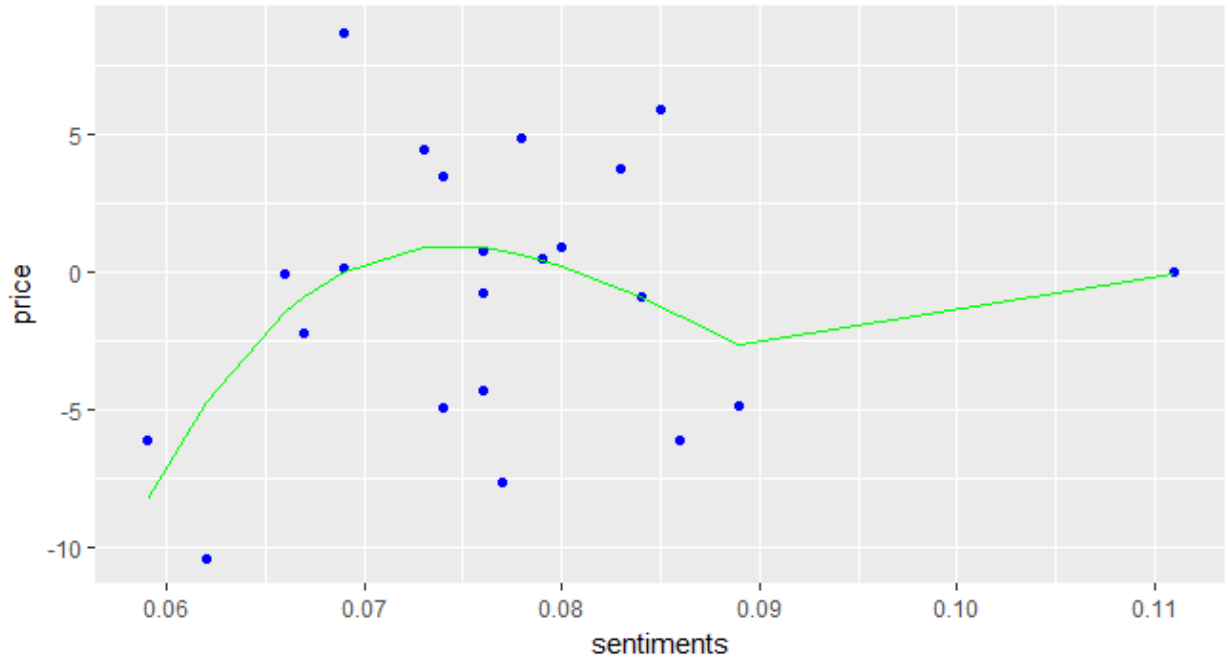

Figure 5.3: Result of testing on Polynomial Linear Regression

4.4 Support Vector Regression

The result of this algorithm implementation is somewhat similar to Polynomial Linear Regression.

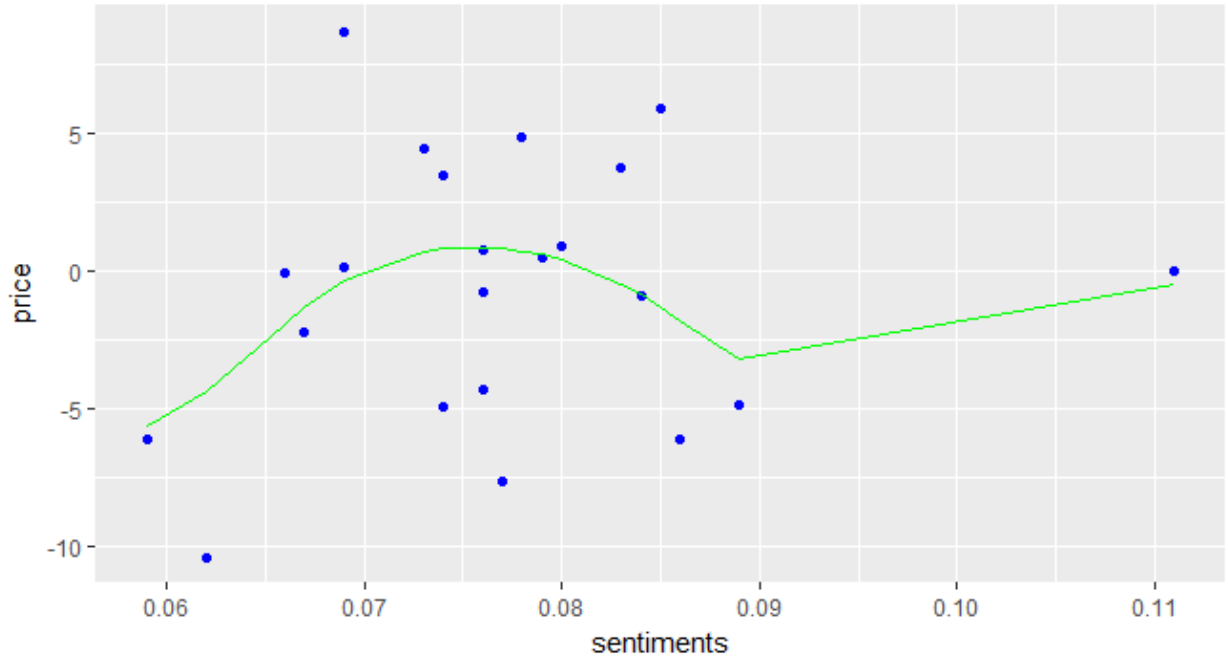

Figure 5.4: Result of Polynomial Linear Regression Testing

4.5 Decision Tree Regression

This algorithm best works for 10 splits meaning when the decision tree is divided into 10 branches. Following is the result for some of the testing:

TABLE III: Decision Tree Regression Results

Actual Result	0.92	0.53	-0.74	-6.09	8.74	-0.02	3.48	0.00
Predicted Result	2.51	2.51833	-2.21	-3.6466	2.79	-5.5	-2.21	-3.6466

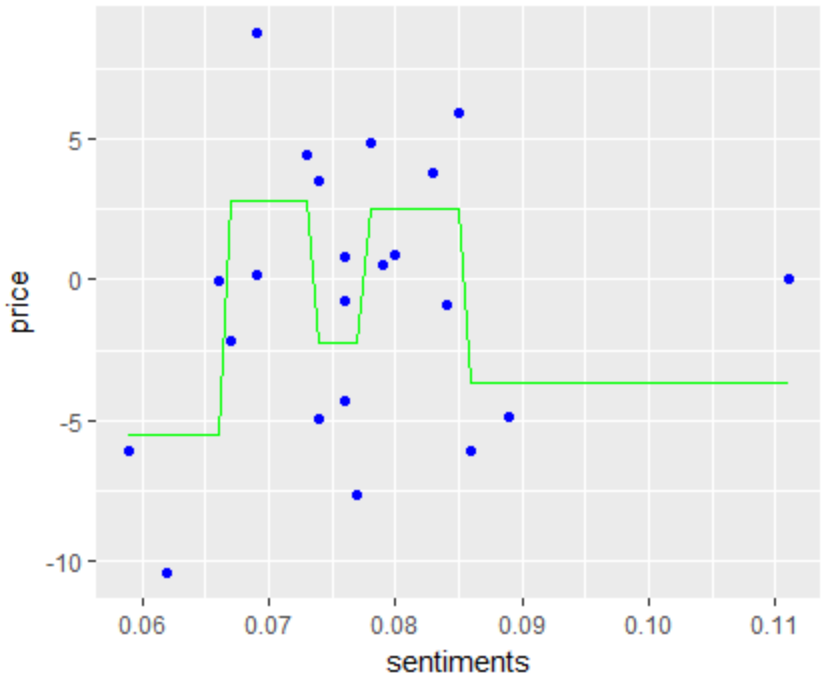

Figure 5.5: Result of Decision Tree Algorithm[27]

4.6 Random Forest Regression

TABLE IV: Random Forest Regression (for 4 Trees) Results

Actual Result	0.92	0.53	-0.74	-6.09	8.74	-0.02	3.48	0.00
Predicted Result	2.02	2.0205	-2.79	2.88	2.66	1.64	2.28	-2.3651

TABLE V: Random Forest Regression (for 1 Tree) Results

Actual Result	0.92	0.53	-0.74	-6.09	8.74	-0.02	3.48	0.00
Predicted Result	3.886	3.886	-1.532	3.886	-0.02	-0.02	-1.532	3.886

Both the results have ⅜ error ration but according to the variance of actual result and predicted result hypothesis with 1 random tree works better.

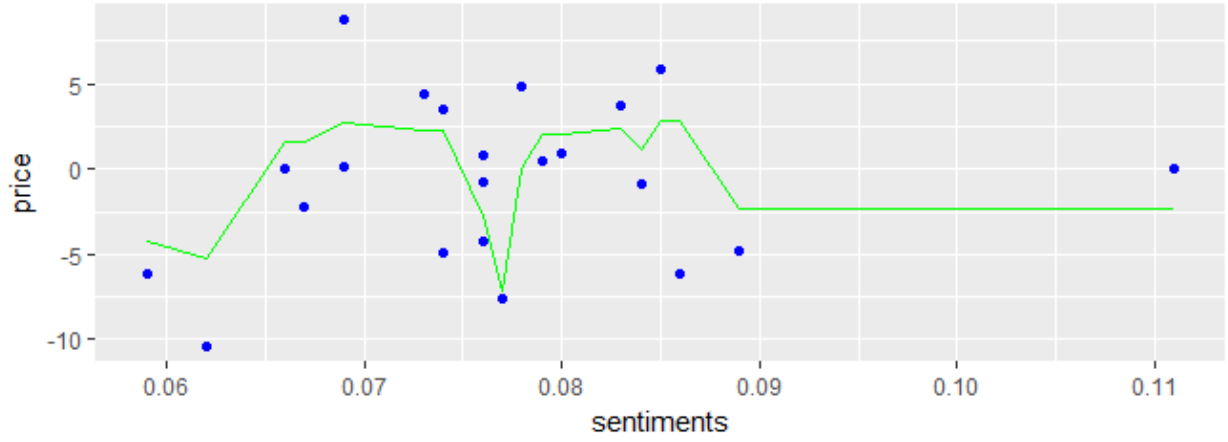

Figure 5.6 (a): Result of Random Tree Regression of 4 decision trees

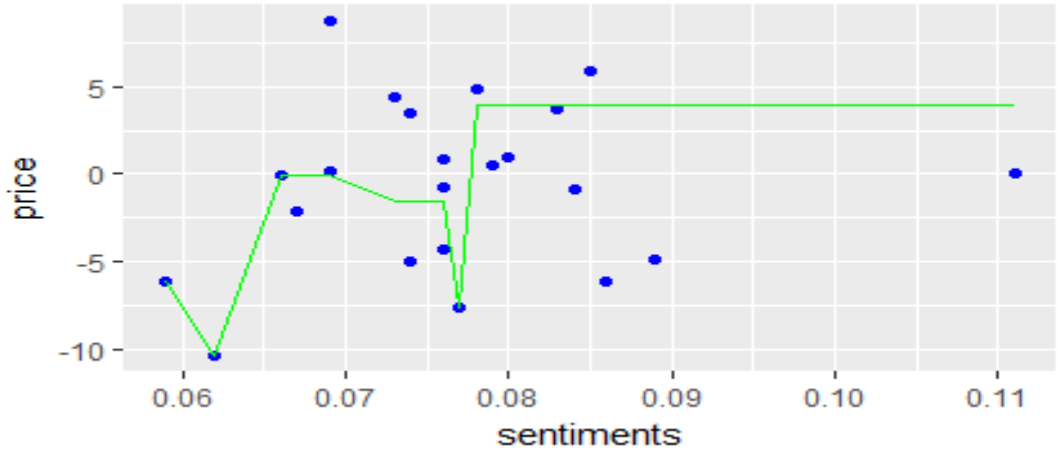

Figure 5.6 (b): Result of Random Tree Regression of 1 decision tree

4.7 Overall Deduction

So, after all the result the conclusion is that with the regression algorithm it cannot be predicted how much the price of cryptocurrency changes. But with the promising result of Multiple Linear Regression, Decision Tree Regression and Random Tree Regression it can be

predicted that weather the price increases or decreases. For Decision Tree Regression the accuracy is 75% which is the best one and also very promising.

4.8 Addressing the problem statements

4.8.1 Statement 1

Yes, we have successfully used our analysis tool ‘R -Studio’ to pull tweets from twitter using their API and extract them into categories such as ‘anger’, ‘fear’, ‘joy’, ‘surprise’, ‘anticipation’, ‘disgust’, ‘sadness’, and ‘trust’, in addition to traditional ‘positive’ and ‘negative’ scores.

We have managed to pre-process the tweets, filter out noise, fill in important missing keywords, correct spellings, read emoticons, and get rid of redundancy as much as possible to acquire scores for each category that composes a complete sentiment of a tweet. To analyze the algorithm for accuracy we have manually identified common and simpler topics to pull out from twitter, scored the resulting tweets using our algorithm into the above-mentioned categories and tested them on 10 selected volunteers, mostly reliable writers and students of the English literature faculty. They were asked to compare the tweets and score them accordingly as per our algorithm. While every judgement is certain to vary, we were more than happy to discover the average sentiment scores provided by our machine and by our human testers only varied to a level that is acceptable. We are also aware of the complexity of our research topic, and understand that the accuracy obtained over simpler topics may not be the same for tweets with keywords such as cryptocurrency, ICO, blockchain etc., since it is obviously, fairly more difficult to categorize the sentiment of tweets regarding Bitcoin than that of a football game, nevertheless we have found out at the conclusion that differentiating a sentiment as ‘fear’ rather than ‘surprise’ is way more significant for our research when trained with a very big dataset rather than identifying by what score they differ in the long run. This is because in the most accurate regression analysis that got us results, we are far more concerned with the variable such as ‘fear’ rather than ‘fear leads surprise by a score of 20’. To simplify- 10 tweets like “I hate bitcoin” serves our purpose as much as 10 tweets like “I wish bitcoin inventors rot in the most

fiery place in hell" because, given the volatility, at the end of the day predicting a general fall in the price the following day is commendable, let alone by what extent.

4.8.2 Statement 2

We have established and compared prediction models for forecasting using machine learning in our work. But to classify it as ideal would be an overstatement, at least at this point, because we haven't been able to test these models over a very long period of time. Since, withstanding the test of time as the cryptocurrency market progresses would be the modest cursor for the model, we can only leave it to future works and implementation to conclude the actual impact of it in the long run.

4.8.3 Statement 3

As much as we can say that our model cannot accurately predict as to what extent the price of Bitcoin may change with change of sentiments, we can re-assure that that it can precisely state whether the price is likely to take a dip or a rise in the immediate future based on results from our work. However, there might be many arguments stating this is not enough in the current volatile market, and we don't disagree, but the whole point of the research was never to pinpoint predictions rather figure out a general trend in light of the various sentiments that come out of social media. Nevertheless, in the short run, an individual can take our insights as educated information and steer investment to a profit margin of around 30 percent.

CHAPTER 05

Conclusion and Future Work

5.1 Conclusion

This thesis studied if we could sub-categorize tweets into precise sentiments and use their scores over a collected interval of time to correlate with the change in price of cryptocurrency, Bitcoin, to be more precise.

Machine learning techniques have been implemented on the gathered sentiments, and quantitative analysis shows fluctuations of accuracy but good precision of judgements as to whether prices of Bitcoin will rise or fall in the short run.

The model best works for Decision Tree Regression with our tested data set with an accuracy of about 75% and lies somewhere in between 30% to 50% with other regression and used techniques. There is also moderate level of correlation between twitter sentiments and price change.

5.2 Future Work

Future work in the research area will begin with acquiring more focused and precise sentiment values to analyse correlation. The more the accurate the sentiments will be, the more certainty in forecasting will be acquired. Additionally, implementing deep learning and neural network algorithms such as RNN, LSTM[17] pose as an exciting prospect to compare and build on our result. Moreover, comparisons can be done using Wavelets[18], and appended to our work to figure out the change in performance. Alternatively, the number dataset used to train our model can be increased and many other different datasets processed with different techniques can be implemented to check for improvement and performance in the long run.

REFERENCES

[1] "Big Data Analytics | IBM Analytics", *Ibm.com*, 2018. [Online]. Available: https://www.ibm.com/analytics/hadoop/big-data-analytics. [Accessed: 01- Apr- 2018].

[2] [Online] https://socialpilot.co/blog/125- amazing-social- media-statistics- know-2016/.

[3] Selfhout M, Burk W, Branje S, et al. Emerging late adolescent friendship networks and Big Five personality traits: a social network approach[J]. Journal of Personality, 2010, 78(2):509C538.

[4] Liu, B.: Sentiment Analysis and Opinion Mining. AAAI-2011, San Francisco, USA, 2011.

[5] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

[6] [Online] http://www.divaportal.org/smash/get/diva2:1110776/FULLTEXT01.pdf

[7] Hong Kee Sul, Alan R Dennis, and Lingyao Ivy Yuan. Trading on twitter: Using social media sentiment to predict stock returns. Decision Sciences, 2016

[8] [Online] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161197

[9] Indurkhya N., Damerau F.J. (2010). Handbook of Natural Language Processing, Chapman & Hall/CRC, Broken Sound Parkway NW, UK. (p. 627-659)

[10] Cambria E., Havasi C. & Hussain A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference. 2-5

[11] Elliott, C. D. 1992. The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System. Ph.D. Dissertation, Northwestern University, Evanston.

[12] Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2):165–210.

[13] Rao, D., and Ravichandran, D. 2009. Semi-supervised polarity lexicon induction. In Proceedings of EACL, 675–682.

[14] Turney, P., and Littman, M. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4):315–346.

[15] O. Poyser. “Exploring the determinants of bitcoin’s price: an application of Bayesian Structural Time Series.” June, 2017

[16] S. Athey. et al. “Bitcoin: Pricing, Adaption and Usage: Theory and Evidence.” 2016

[17] S. McNally. “Predicting the price of Bitcoin using Machine Learning.” MSc Research Project, School of Compuitng, National College of Ireland, 2016

[18] L. Wang, K. K. Teo, and Z. Lin, \Predicting time series with wavelet packet neural networks," in Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on, vol. 3. IEEE, 2001, pp. 1593{1597.

[19] D. Shah and K. Zhang, \Bayesian regression and bitcoin," in Communication, Con-trol, and Computing (Allerton), 2014 52nd Annual Allerton Conference on. IEEE, 2014, pp. 409{414.

[20] B. G. Malkiel and E. F. Fama, \E cient capital markets: A review of theory and empirical work," The journal of Finance, vol. 25, no. 2, pp. 383{417, 1970.

[21] G. H. Chen, S. Nikolov, and D. Shah, \A latent source model for nonparametric time series classi cation," in Advances in Neural Information Processing Systems, 2013, pp. 1088{1096.

[22] I. Georgoula, D. Pournarakis, C. Bilanakos, D. N. Sotiropoulos, and G. M. Giag-lis, \Using time-series and sentiment analysis to detect the determinants of bitcoin prices," Available at SSRN 2607167, 2015.

[23] M. Matta, I. Lunesu, and M. Marchesi, \Bitcoin spread prediction using social and web search media," Proceedings of DeCAT, 2015.

[24] L. Kristoufek, \What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis," PloS one, vol. 10, no. 4, p. e0123923, 2015.

[25] I. Madan, S. Saluja, and A. Zhao, \Automated bitcoin trading via machine learning algorithms," 2015.

[26] A. Greaves and B. Au, \Using the bitcoin transaction graph to predict the price of bitcoin," 2015.

[27] K. Eremenko, H. de Ponteves, SuperDaraScience Team (2018). “Machine Learning A-Z™: Hands-On Python & R in Data Science.