IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Automated Trading of Cryptocurrency Using Twitter Sentimental Analysis

S.R. Chheda[1*], A.K.Singh[2], P.S. Singh[3], A.S. Bhole[4]

Dept. of Computer Science and Engineering, Ramdeobaba College of Engineering and Management, Nagpur, India

*Corresponding Author: shrustichheda@gmail.com, Tel.: +917507553496

Available online at: www.ijcseonline.org

Accepted: 23/May/2018, Published: 31/May/2018

Abstract— Twitter is one of the most used social networking sites where millions of people give their opinions about various subjects. Thus it can be treated as one of the largest and most updated psychological database. Analysis can be performed on this data to gain valuable insights. The goal of this paper is to study the correlation between public opinion about the Cryptocurrency, Bitcoin and the trajectory of its price graph. The results can later be used to develop a system for algorithmic trading of Bitcoin. This is done by collecting tweets on bitcoin and performing sentimental analysis on it. The tweets are labelled positive or negative. Supervised machine learning algorithms are used to see how sentiments of tweets play a role in bitcoin market movement. A positive sentiment and an increase in the price of bitcoin at the same time will indicate selling as favourable and vice versa in case of a negative polarity of tweets.

Keywords—*Twitter, opinions, sentimental analysis, bitcoin, machine learning algorithms*

I. INTRODUCTION

A cryptocurrency can be viewed as a digital or virtual currency which works as a medium of exchange. Numerous cryptographic techniques are used to secure and verify transactions. They are also used to control the creation of new units of a particular cryptocurrency. Essentially, cryptocurrencies are limited entries in a database that no one can change unless specific conditions are fulfilled. Since its origination in 2009, Bitcoin has received the stature of a digital commodity and its worth is considered comparable to our traditional currency. Considering the fluctuation in exchange rates of cryptocurrency we try to develop a trading strategy that can be applied to a variety of cryptocurrencies. This is done by applying sentimental analysis on twitter data and later classifying this data. With the advent of the information age the ability to identify and categorise sentiments has become increasingly important for businesses and researchers as well [1]. Using Twitter for data is favourable because it is one of the largest and most effective social media networks. Using twitter also means using the earliest and fastest intelligence update that too in a concise format. Also data can be mined from twitter with relative ease.

We use supervised machine learning algorithms like Support vector machine and Naive Bayes to predict the sentiments of tweets using which we predict the prices of Cryptocurrency[1]. The classifiers are trained using two approaches i.e. first, training them directly through texts and other using sentiment analysis APIs which allot sentiments to each tweet i.e. positive or negative sentiments. We then compare the predicted sentiments to find out if the behaviour is positive or negative towards the price of cryptocurrency.

Rest of the paper is organized as follows, Section II contains the related work in the field, Section III contains the methodology adopted for implementation, Section IV discusses the results and analysis, section V concludes the discussions and results obtained, Section VI talks about future work while Section VII is the list of references.

Figure 1: Fluctuations in bitcoin prices

II. RELATED WORK

Twitter

With increasing use of the internet, the way people express their ideas and opinions has changed. Because of its advantages over the traditional media in terms of wider reach, frequency, immediacy, usability and permanence more industries use social media to distribute information [2]. Twitter is a popular micro blogging site which is equipped with various APIs favourable for data mining. Twitter is also termed as the most updated psychological database.

Bitcoin

Bitcoin is one of the many cryptocurrencies available today. It is the most established and discussed cryptocurrency and its value is considered comparable to the traditional currencies in use. It is the first decentralized digital currency, as the system works without an administrator or a central bank. The exchanges of bitcoin are verified for secure transaction by network nodes which use cryptographic techniques. They are recorded in a public distributed ledger called a block chain which records bitcoin transactions. It is implemented as a chain of blocks, each block containing a hash of the previous block up to the genesis block. The reason bitcoin is chosen among other cryptocurrencies is because it is well established and trusted commodity in financial markets.

Sentiment Analysis

Using sentimental analysis for trading of cryptocurrencies is a part of a wide field called behavioural economics which is a method of economic analysis. It uses psychological insights into human behaviour and thus explains economic decision-making. With the increase in machine learning technologies this task is even more simplified. The use of machine learning in trading is relatively new, but fair amount of research has been done on these subjects. Shah et al. in his paper discusses the method of Bayesian regression and its efficacy for predicting price variation of Bitcoin. Sul et al. [3] in the paper titled *Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns* studies the sentiment of nearly thirty five lakh tweets about S&P 500 firms. According to the results the sentiments on social media are reflected in the stock prices of a firm the same day or the next day. Sentimental analysis in itself is a very wide branch. Sentiment analysis can be divided into three levels namely document level, sentence level and phrase level. While at the document level opinion of the whole document is checked, the sentence level is where polarity of sentence is determined. The analysis at entity level gives fine grained analysis of a particular entity [4]. Various researches have been done, initially classifying documents by (Turney, 2002; Pang and Lee, 2004), then performing sentence level classification (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently classification at the phrase level (Wilson et al., 2005; Agarwal et al., 2009). A significant amount of work on sentiment analysis of the Twitter data is conducted by Go et al. (2009), Bermingham and Smeaton (2010) and Pak and Paroubek (2010). Through this project the effect of the sentiments of people on the bitcoin markets is observed by referring to the previous studies in this field.

Objectives

1. Create training and testing data set for the learning algorithms.
2. To study the correlation of public sentiment with the rise or fall in the price of cryptocurrency.
3. To determine digital currency (Bitcoin) market movement with the Twitter data set, text classification and sentiment analysis algorithms.

III. METHODOLOGY

Data Collection and Pre-processing

The process of data collection is started using Tweepy which is an open source Python library, by assigning the search query as 'bitcoin'. We set the number of tweets to be collected to 1000 in a single run. The tweets are collected in a text file in a specific format as follows: Author|| Tweet || Timestamp. Noisy datasets are obtained from raw tweets scraped from twitter. First basic text pre-processing steps are followed. Tweets have certain special characteristics such as re-tweets, user mentions etc. which have to be suitably extracted. Therefore, raw twitter data should be normalized to form a dataset which may be simply learned by numerous classifiers. We have applied an intensive variety of pre-processing steps to standardize the dataset and scale back its size. We convert the tweets to lowercase, replaced two or more dots (.) with space, striped spaces and quotes (" and ') from the ends of tweet and replaced two or more spaces with a single space. We handle unique twitter features as mentioned in Table1.

After applying tweet stage pre-processing, we process individual phrases of tweets as follows.

- Strip any punctuation [:;?()!,"'.] from the phrase.
- Convert two or more letter repetitions to two letters. People often send a tweet like this is toooo good, adding more than one character to emphasise on certain phrases. We handle such tweets by converting them to a standardized form like this is too good.
- Remove - and '. This is done to handle words like T-rex and her's by converting them to the more general form Trex and hers.

- Stop words are subsequently removed from tweets based on membership in the "stop words" corpus of the Natural Language Toolkit.[5]
- Check the validity of the words and accept the word only if it is valid. A valid word is defined as a word which begins with an alphabet with alphabets being the successive characters, numbers or one of dot (.) and underscore (_).

Table 1. Data Pre-processing Stage

Feature	Use	Pattern Matching	Replaced with
URL	**Users often share hyperlink to other web pages in their tweets**	**((www\.[\S]+)|(https?://[\S]+))**	**URL**
User Mention	**Users often mention other users in their tweets by @handle**	**@[\S]+**	**USER_MENTION**
Hashtag	**Hashtags are unspaced phrases prefixed by the hash symbol (#) which is frequently used by users to mention a trending topic on twitter.**	**#(\S+)**	**#hello is replaced by hello**
Re-tweet	**Re-tweets are tweets which have already been sent by someone else and are shared by other users. Re-tweets begin with the letters RT**	**\brt\b**	**Remove RT**

Sentimental Analysis

For opinion mining from the tweets we use a python library Text Blob. It divides the Natural Language into its various grammatical parts such as parts-of-speech, noun phrases etc. It performs tasks such as tagging, classification and sentimental analysis. The Text Blob API assigns Polarity scores which is a float within the range [-1, 1]. While processing we remove the neutral value i.e. polarity as these tweets are considered objective in nature with no opinions or neutral opinions. We run our text files obtained after pre-processing through the sentimental analysis script. The output is saved in CSV files which contain the polarity label along with the tweets and a unique id.

Feature Extraction

We use Unigrams and Bigrams as the two features for prediction. Frequency distribution of Unigrams and Bigrams is done to obtain the feature vectors which are later fed to the classifiers.

Unigrams and Bigrams

The tweets need to be tokenized first before we process them. In the fields of computational linguistics and probability, an n-**gram** is a contiguous sequence of *n* items from a given sample of text or speech. The *n*-grams typically are collected from a text or speech corpus [6]. When the value of n is 1 it is called Unigram and when n is 2, it is called Bigram. Using python scripts we find out the frequency distribution of unique unigrams and bigrams. After removing the noise words we find close to 15000 unigrams and 90000 bigrams. These are used to create the testing vocabulary. These are converted to feature vectors by representing them in a sparse vector representation. We create the vectors on feature type frequency. Thus a positive value is allotted at the indices of unigrams or bigram which gives their frequency count. This is done for the entire training set. The frequency is increased or scaled by the inverse-document-frequency to give more weightage to important words.

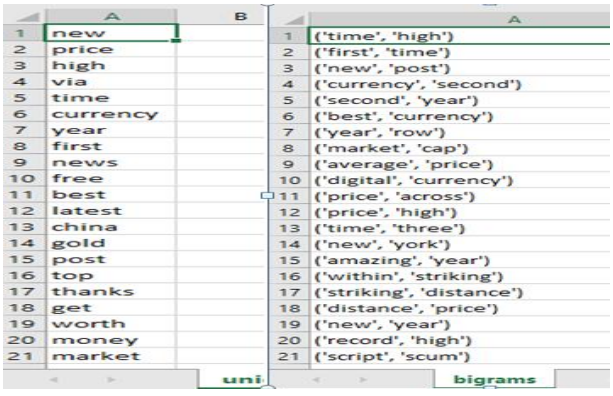

Figure 2: A sample of unigram and bigram CSV file

Testing the accuracy of training set

We predict the sentiment of each tweet using an opinion dictionary that contain both positive and negative. We count the number of positive words and negative words in each tweet matching the words from this dictionary and if number of positive words are greater than negative words then positive sentiment is assigned to tweet and negative sentiment is assigned if vice versa. If both negative and positive sentiments are same the positive sentiment is assigned. The predicted value is compared with value assigned by API to get the accuracy of training set.

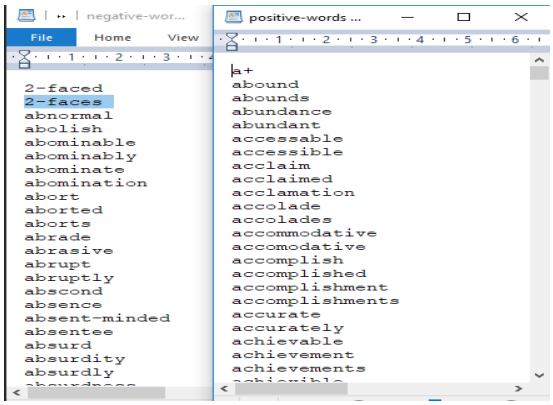

Figure 3: List of positive and negative words used for classification

Figure 4: Testing accuracy of training data

Classifiers

Bayesian network classifiers are a popular supervised classification paradigm. A well-known Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier based on the Bayes' theorem.[7] Bayes theorem provides a way of calculating posterior probability

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)P(c)}{P(x)} \quad (1)$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Multinomial Naive Bayes is used from Scikit-learn [8] for naive Bayes classification. Multinomial Naïve Bayes is used for discrete counts. We ran the experiment using frequency as feature type. We noticed that addition of bigram in feature vector increases the accuracy. The best accuracy obtained is 88.72% which was using both unigram and bigram.

"Support Vector Machine" (SVM) is non probabilistic algorithm which is used to separate data linearly and nonlinearly [7]. Here each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well. On running the experiment on our data the accuracy obtained is 89.02%

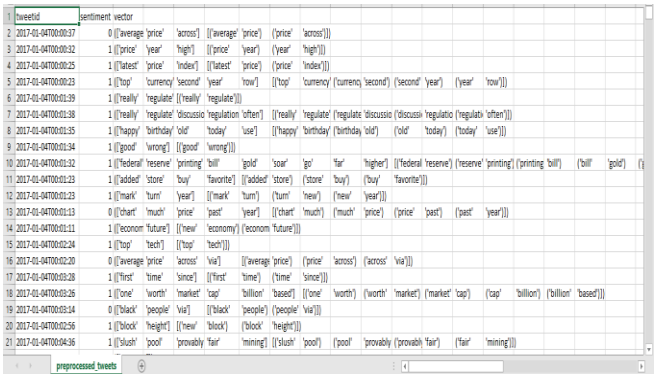

Figure 5: Feature vector given as input to classifiers

IV. RESULTS AND DISCUSSION

We run the classification algorithms on the training data set. Out of this 10% of data is used for testing. The results of accuracy for Naïve Bayes and SVM are summarized in the table 2.

Table 2. Accuracies of the two classifiers

Classifier	Unigram	Unigram and Bigram
Naïve Bayes	84.0405	88.7283
SVM	86.1965	89.0173

We process the tweets of 4th and 6th January 2017. The results are stored and compared to the prices of bitcoin for these two days. The tweets are classified and the sentiment labels are aggregated. Negative sentiment matches with a decrease in price while a hike in prices can be observed in case of positive sentiment in tweets.
The sample output of accuracies considering the day-wise collected tweets on 04-01-2017 as a training set .and 06-01-2017 as testing data set is shown in table 3.

Table 3. Accuracies obtained for particular dates

Feature vectors	6 Jan 2017
Unigrams	82.1251%
Bigrams	84.0405%

When positive sentiments are more in the data then the value is expected to increase and thus trading can be done accordingly. Vice-versa happens when negative sentiments are more.

Further to verify our results, they were tested on live data. We process the tweets of 18th and 19th April 2018. The results are stored and compared to the prices of bitcoin for those two days. The tweets classified as negative matched with a decrease in price while a hike in prices is observed in case of positive tweets.

V. CONCLUSION

We get the predicted value from the classifiers. We sum up the number of positive sentiments and negative sentiments to find the final behaviour. If the number of positive sentiments are more than number of negative sentiments then price of cryptocurrency is predicted to increase else if the number of negative sentiments are more the price is predicted to decrease. If the numbers are same then the prediction says no change in price. Thus we can conclude that sentiments and opinions of people reflect in the trajectory of price graph of Bitcoin.

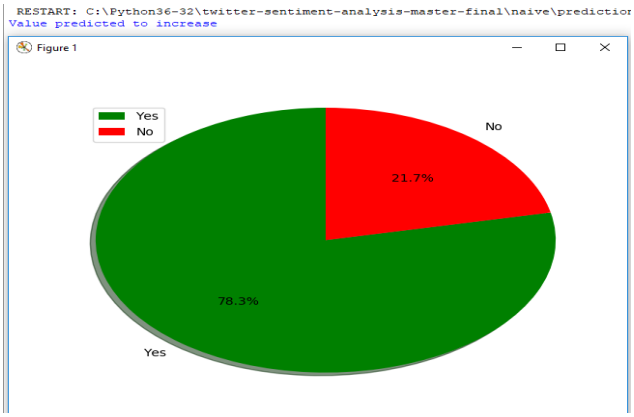

Figure 6: Predicting rise or fall in prices based on sentiments

VI. FUTURE WORK

1. The range of sentiments could be increased to give more accurate results. For example the range could be taken from -2 to 2 where -2 is most negative and 2 is most positive. Thus intensity of polarity of tweets can be taken into account.
2. We could provide sentiments based on domain of the person. For example the tweets of a financial market analyst could be given more value.
3. During pre-processing we discard symbols like commas, full stop and other special characters. These symbols are helpful as they can be used to assign sentiments to tweets. But more complex processing is required to deal with the same.

VII. REFERENCES

[1] S.Colianni, S.Rosales, M.Signorotti , "*Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis*", Department of Computer Science Project, Stanford University, 2015.
[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. "*Finding high-quality content in social media*". in Proceedings of the international conference on Web search and web data mining. 2008. ACM.
[3] H.K. Sul, A.R. Dennis, and L.I. Yuan, "*Trading on twitter: Using social media sentiment to predict stock returns*" Decision Sciences, 2016.
[4] C. Nanda , M. Dua, "A Survey on Sentiment Analysis", International Journal of Scientific Research in Computer Science and Engineering (0975 – 8887) Vol.5(2), April(2018), E-ISSN: 2320-7639
[5] S .Bird, E. Klein,E. Loper."*Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit*" ,O'Reilly Media Inc, USA,pp 39-250,2009.

[6] Mamatha M, Thriveni J , K.R.Venugopal, "*Techniques of Sentiment Classification, Emotion Detection, Feature Extraction and Sentiment Analysis* ", International Journal of Computer Sciences and Engineering Vol.6(1), Jan 2018, E-ISSN: 2347-2693

[7] B.M. Jadav ,V.B. Vaghela, "*Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis*", International Journal of Computer Applications (0975 – 8887) Volume 146 – No.13, July 2016.

[8] R.Garreta , G.Moncecchi , "*Learning Scikit-learn: Machine Learning in Python" PACKT publishing* , India, pp 24-96 , 2013

[9] E.Stenqvist, J. Lonno, "*Predicting Bitcoin price fluctuation with Twitter sentiment analysis*", DEGREE PROJECT IN TECHNOLOGY, FIRST CYCLE

Authors Profile

Ms Shrusti Chheda is currently pursuing Bachelors of Engineering in Computer Science from Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India.

.

Mr Aman Singh is currently pursuing Bachelors of Engineering in Computer Science from Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Ms Priti Science is currently pursuing Bachelors of Engineering in Computer Science from Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India.

Mr Aditya Bhole is currently pursuing Bachelors of Engineering in Computer Science from Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India.