

Evaluating Sentiment Classifiers for Bitcoin Tweets in Price Prediction Task

Ahmed M. Balfagih, Vlado Keselj,
Faculty of Computer Science, *Dalhousie University*

Abstract— Bitcoin alongside other cryptocurrencies became one of the largest trends recently, due to its redefinition of the concept of money, and its price fluctuation. Especially on the social media, people keep discussing Bitcoin topics, consulting, and advising about cryptocurrency trading. This paper explores the relationship between Twitter feed on Bitcoin and sentiment analysis of it, comparing and evaluating different data mining classifiers and deep learning methods that might help in better sentiment classification of Bitcoin tweets, the study uses different language modeling approaches, such as tweet embedding and N-Gram modeling. We also evaluate the quality of automated sentiment classification in comparison to manually assigned sentiment labeling. The results show that the manual approach gives significantly better results in some datasets, and superior performance of MLP, WiSARD and decision tree methods. On the other hand, R-Auto Tweets Sentiment (RATS) gives more stable performance overall datasets. Using time-series, we found partial correlation between Bitcoin price fluctuation and sentiment class accuracy fluctuations using different machine learning algorithms.

Index Terms— *Bitcoin, Cryptocurrency, Twitter, Social media, Sentiment, Prediction, Classification, Machine learning, Text mining, Natural language processing, Time-Series Data.*

1 INTRODUCTION

Cryptocurrency is a representation of digital property. More precisely, it is a program written in a specific programming language and using universal encryption techniques that make it impossible to penetrate and manipulate. Cryptocurrency is a term used in the science of encryption for security purposes to protect virtual transactions and control the establishment of new units. Therefore, it is difficult to fake that currency, which is a virtual exchange medium (electronically), and a branch of the alternative currency [1]. Moreover, cryptocurrency is used to denote all these applications using Blockchain, whether these applications represent a digital currency or represent other things such as smart contracts and others [2]. Bitcoin was the first cryptocurrency created in 2009 and since then other currencies have been created for its competitors such as Litecoin, Nemcoin, and others. Cryptocurrencies have many advantages, firstly that they do not have a central authority, unlike other electronic currency systems, such as PayPal. secondly this is where the public record account, which usually called consensus, where the virtual transactions are recorded in an integrated manner. Cryptocurrencies are easily used to transfer the balance between the parties in transactions. These transfers go around the use of public and private encryption keys for security purposes. The transfer of balances ends with lower processing fees, allowing users to avoid sharp shipping fees from most banks and financial institutions. In central banks, economic systems such as the Federal Reserve, the public sector and governments are control the currency vaults by printing the fiat money. However, companies and governments cannot produce encrypted currency units or in other words yet there are no bonds provided to other

companies, banks and corporations entities where the value of assets measured in a decoded currency is decentralized and the currency is created by a complete technical and fundamental system where it is created by A known or no-known group or character such as the Bitcoin inventor.

The main disadvantage of cryptocurrencies is that they are virtual. They do not have a central store. A user's budget for an encrypted digital currency can be completely erased due to a computer failure if there are no backups of their holdings. Anonymous financial transactions in encrypted currency are, of course, perfectly suited to egregious activities such as tax evasion and money laundering. Therefore, encrypted currencies are not centralized because of security in theory of manipulation or government intervention. While there are hundreds of cryptocurrency features, most of which are derived from one or two protocols: proof of work and verification of the test, all these encrypted currencies are reserved by the mineralization of encrypted coins and are equipped specifically with the specific computers or currency mining device (ASIC Miner) to participate in the interaction and procedures of transactions. Cryptocurrencies became a popular trend these days and gained a lot of people interest especially whom are interested in investing their money or looking for an alternative solution for speculation to increase their income. Bitcoin proves that this field is fertile for investing, given that its price has been doubled at least 50,000 times for seven years. Mainly, coin value relies on demand and supply, and maybe other factors that related to economics news or different politics.

2 BITCOIN PRICE AND TWEETS

Convictions of people about cryptocurrencies have been changing, and cryptocurrencies are continuing to prove their existence as an alternative money. Investing and speculation are two of the most attractive topics for people who are interested in increasing their income. Similarly, cryptocurrency market has some features in common with markets such as stock market, foreign currency market (forex), and other assets markets like crude oil, gold, and different valuable metals. Many factors may affect the price of different coins ups and down, related to the volume of demand and supply, and some other economic and political news and events. It is very important for investors and speculators to have such tools that predict the increase and decrease in the cryptocurrencies prices and suggest to them what currency is better to invest in. It is useful to take advantage of social media and trends about cryptocurrency and investigate if there is a strong correlation between people's posts and coins prices changes.

Therefore, the research investigates these aspects:

- Is there a correlation between Twitter sentiment and BTC fluctuation?
- Can machine learning model based on polarity sentiment accuracy be used for Bitcoin price prediction?
- Is automated sentiment classification better than manual labeling?

The paper has positive prospective impact on the market of the cryptocurrency, that affects both the investor and the currency itself. If the investor was enabled to get rid of fear of losing his money by having a tool that help him to predict prices and advise him as a consulter, he will be encouraged to invest more, therefore, he will have higher profit potential. This may also be reflected on the currency, that people would be recommended to increase their demand for it. The more purchases of such a coin, the more the increase in its price and market capital.

3 RELATED WORK

Stock market is like cryptocurrency market. A stock market is a platform to buy and sell stocks, and it driven by supply and demand, so the cryptocurrency market stock. The main players in the stock market are the exchanges. Exchanges are where the sellers are matched with buyers to both facilitate trading and to help set the price of the shares. Stock exchange is the efficient way to create a favorable climate for an active and growing for new issues. Statistical algorithms and time series analysis have been used to predict the stock market. ABIRAMI, R. and VIJAYA, M.S [4], employed machine learning technology in computational finance since machine learning deals with techniques that allow computers automatically learn to make accurate predictions based on past observations. They used WEKA environment. [5]

for training the dataset using linear regression, and LIBSVM tool for support vector regression.

In another study, PAGOLU et al. used microblogging social media to predict stocks price, because it is perfectly representing the public sentiment and opinion about current events [6]. Stock market prediction based on public sentiments expressed on Twitter has been an intriguing field of research. The thesis of this work is to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company. Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper has employed two different textual representations, Word2vec and Ngram, for analyzing the public sentiments in tweets. They applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyze the correlation between stock market movements of a company and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company would encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. The researchers concluded that the paper is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets.

MATA et al. explored in case the rise of the Bitcoin's cost is related to the volume of tweets or Web Search media outcomes. They compared patterns of cost with Google Trend Data, volume of tweets and especially with positive tweets [7]. Tweets are accessible and are effectively imported from Twitter Application Programming Interface (API). Composing the hashtag #Bitcoin or mention @Bitcoin, then they assemble all tweets that specified the analyzed subject. Beside that they imported data from Google trend, and used DataStore (back-end database engine, utilizing MySQL as RDBMS), SentiStrenght tool, and Java Module. They collected more than 1,900,000 tweets, then analyzed them and identify the positive tweets and negative ones. Using cross-correlation computation, they found similarity between Bitcoin Price with the number of tweets in same time, between Bitcoin price with the positive tweets and Bitcoin price with Google trend.

STENQVIST and LÖNNÖ, 2017 [8] studied if sentiment analysis on Twitter data can help to indicate to Bitcoin price fluctuation. They applied a naïve prediction model based on the amount of sentiment fluctuations over time-series, using intervals from five minutes up to four hours and from one to four shifts. By analyzing over two million tweets related to bitcoin within one month (day-to-day analyzing), they found that the most accurate aggregated time was one-hour interval indicating a change after four hours. They mentioned that applying machine learning to the study may has a correlation further than the number of tweets.

4 PRELIMINARY SYSTEM ANALYSIS AND DATA PREPROCESSING

4.1 Social Media Data Selection

In this study, dataset has been chosen from the social network Twitter based on the following reasons:

1. The availability of huge number of training tweets, considering that Twitter users had already passed 300 million users, tweeting every second, on average, around 6,000 tweets, which corresponds to over 350,000 tweets sent per minute, 500 million per day and around 200 billion per year. Therefore, Bitcoin's traders can find live feeds regarding Bitcoin.
2. The accessibility to collect Bitcoin tweets using a Twitter API tool.
3. Twitter is a microblogging social media platform, therefore, analyzing short blogs or 'tweets' is easier than analyzing articles with big corpus.

4.2 System Architecture

The general form of the proposed system consists of three major parts: (1) Data pre-processing part which includes data importing from Twitter, data cleaning, data modeling process and tweet's sentiment process. (2) Data mining modeling which includes multiple classification experiments and tests. (3) and finding events and correlations using time series. See **Figure 1**.



Figure 1 The proposed system architecture.

The data mining tools that have been chosen are Weka and R, due to the abundance of research studies that are like this study and used the same tool. Weka will be used to analyze the dataset information and to design the appropriate data mining model. R tool will be used too in the data preprocessing phase.

4.3 Data Preparation

Five lists of historical tweets that related to Bitcoin have been collected to be applied in this study. Two of them are considered short datasets (less than 12 hours), and the other three are 24-hour tweets' datasets. The short two

datasets are tweets were collected from a period that the price of Bitcoin was skyrocketing (December the 12th, 2017 – five hours period), and tweets were collected from a period that the price of Bitcoin had a significant drop (March the 23rd, 2018 – two hours and half period).

The variety of datasets relying on the date and period and number of tweets has been considered. All datasets were found in *Kaggle.com* and gripped using API tool from Twitter [3]. **Table 1** shows the details of historical tweets, and the number of tweets after removing irrelevant tweets. Some features such as tweet URL, or Twitter user were removed because it is not relevant to the tweet sentiment analysis, thus, it is better to focus on the tweet text, which includes hashtags as well. The data cleaning process applied using a code implemented by R language. All punctuations were removed, and all letters were transformed to uppercase. All irrelevant tweets has been removed to become a total of 282K tweets.

Sentiment Class	R Auto-Sentiment analyzed tweet example
Positive	RT DOMENCLATURE KEEP YOUR EYES ON THE BALL HAPPYHANUKKAH GROWTH TUESDAY THOUGHTS STARTUP BIGDATA DOMAINING TECH BITCOIN CRYPTOCU EBAY TAKING BTC BITCOIN FOR PAYMENTSWELL IT WORKED OR VERY WELL FOR OVERSTOCK IF THEY WANT THEIR SALES TO
Neutral	RT SALGORITHIMS FOXBUSINESS POTUS DIGIBYTE IS 40X FAST THAN BITCOIN AND MORE SECURE NOT TO MENTION WAY CHEAPER TO SEND DGB SEE FOR TCOT BITCOIN AND CRYPTO CURRENCIES WHAT YOU SHOULD KNOW VIA YOUTUBE
Negative	RT ROGERPARKEY CRITICISMS AGAINST BITCOIN ARE OFTEN MISLEADING EYS ANGUS CHAMPION SHARES WHAT BITCOIN AND BLOCKCHAIN MEANS FOR THE MT GOX CREDITORS WANT BITCOIN EXCHANGE TAKEN OUT OF BANKRUPTCY BITCOIN CRYPTO

Table 1 The total of tweets are more than 282K after removing irrelative tweets

4.4 Tweets Sentiment

Two approaches of tweet sentiment were applied: Manual sentiment and auto-tweet sentiment using R language or acronym (R-Auto Tweet Sentiment/ RATS) [9]. In the manual sentiment approach, each tweet has been read and decide to classify it into three different classes: positive, negative and neutral. This is a time-consuming process in comparison to the RATS, that can be used by TwitterR package and `get_nrc_sentiment(dataframe)` function. The RATS generate values ranged between "3" and "-3", and the values that is between "1" and "-1" were considered neutral. This will form two different versions of each dataset to be tested, and to discover which one sentiment approach is more reliable and efficient in tweets sentiments. **Table 2** shows an example of how RATS classify tweets to positive, neutral and negative sentiment.

Datasets	Date	Period	# of Tweets
Dataset1	5 Aug 2017	24 Hours	14,462 Tweets
Dataset2	12 Dec 2017	5 Hours	9,647 Tweets
Dataset3	23 Mar 2018	2.5 Hours	10,207 Tweets
Dataset4	12 Jul 2018	24 Hours	166,444 Tweets
Dataset5	15 Mar 2019	24 Hours	81,610 Tweets
Total			282,370 Tweets

Table 2 Example of tweets classified into polarity sentiment using RATS after data preparation process

4.5 Language Modeling Approaches

1. **Tweet Embedding:** while word embedding is one of recognized approaches of modeling words into numbers to be used in text mining and deep learning. Weka has a tweet embedding function in a package named “Affective Tweet” that does same role on the whole tweet.

The function `TweetToEmbeddingsFeatureVector` calculates a tweet-level feature representation using pre-trained word embeddings. A dummy word-embedding formed by zeroes is used for word with no corresponding embedding, then average word-embedding is calculated and generate 100 embedding values for each tweet.

2. **N-Gram** is another data modeling approach that is commonly used in natural language processing studies. In this process a unique unigram/bigram vocabulary is built to be used as features for each tweet. Using Weka, `StringToWordVector` function is applied to generate a unigram/bigram keyword. This function converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings. **Ngram** tokenizer is applied, and elimination of terms that do not appear at least 5 times in a dataset. Moreover, *idf* and *tf* transforms are applied in this process.

3. **Tweet Embedding and N-Gram** is a combination of generated features in one data model. This is to observe if using both approaches together has a positive impact on classification accuracy.

Therefore, 30 different datasets are formed to be tested in the data mining process, 15 with manual sentiment and the rest with R-sentiment. These different features datasets will be applied to machine learning algorithm to evaluate the best form of modeled features. **Table 3** shows all forms of the dataset after tweet embedding and N-Gram for with number of tweets for positive, negative and neutral tweets.







Datasets			Positive Sentiment		Negative Sentiment		Neutral Sentiment		Total of Tweets
Name	Model	Features							
Dataset 1 5Aug17	Tweet Emb.	102	4,481	4,056	1,739	1,917	8,242	8,489	14,462
	N-Gram	471							
	Twe. Emb. + N-Gram	571							
Dataset 2 12Dec17	Tweet Emb.	102	1,523	3,139	1,503	1,427	6,621	5,081	9,647
	N-Gram	1147							
	Twe. Emb. + N-Gram	1247							
Dataset 3 23Mar18	Tweet Emb.	102	4,721	5,075	1,249	1,130	4,237	4,002	10,207
	N-Gram	1245							
	Twe. Emb. + N-Gram	1345							
Dataset 4 12Jul18	Tweet Emb.	102	46,008	36,268	25,064	10,757	95,372	119,419	166,444
	N-Gram	443							
	Twe. Emb. + N-Gram	543							
Dataset 5 15Mar19	Tweet Emb.	102	24,245	17,009	13,174	5,912	44,190	58,689	81,610
	N-Gram	441							
	Twe. Emb. + N-Gram	541							
30 forms of datasets									282,370

Table 3 All formed dataset after generating features of Tweet-Embedding and N-Gram, with numbers of tweets for each sentiment

5 MACHINE LEARNING METHODS

5.1 Classification Methods

After finishing the preprocessing phase and having different remodeled datasets, it is obvious that we need to apply a supervised data mining technique to predict sentiments. The plan is to test the most common classification algorithms and compare the accuracy of prediction to each other. Using Weka, the algorithms were chosen relying various classification methods' concepts: lazy learner method, Bayesian method, and decision tree method. Five different classification algorithms have been chosen: K-Nearest Neighbor (KNN) [12], Bayesian Network (BN) [13], Naïve Bayes (NB) [14], C4.5 Decision Tree Algorithm (J48) [15], Random Forest Decision Tree (RF) [16].

5.2 Deep Learning Methods

Other neural network methods are applied to figure out its accuracy and efficiency, these methods are available in Weka workbench using the deep learning package Weka-Deeplearning4j, which is a Java library developed to incorporate the modern techniques of deep learning into Weka. The deep learning algorithms that applied are: Multi-layers Perception method using `D14jMlpClassifier` function with two dense layers (MLP) [17], Radial Basis Function Network (RBFN) [18], and Weightless Neural Network (WiSARD) [19].

6 TIME-SERIES

To find correlation between Bitcoin price fluctuation and the Twitter sentiment, different time series frequencies are applied on the five main datasets. The short datasets are divided into one interval that is proper for the number of tweets that contain (fifteen minutes for dataset # 2 and five minutes for dataset # 3), while the remaining datasets are divided into three intervals ranging from 30 minutes up to 120 minutes. Each interval will measure the fluctuation of positive sentiment accuracy (the true positive value) and the negative class accuracy, then try to find the correlation with the Bitcoin price change. Two interval shifts will be applied to predict matching event, single shift and double shift.

Matching event for the positive class sentiment will be any rise or descent for the true positive value of the positive class in one interval with same fluctuation of the Bitcoin price in the next shift. While matching event for the negative class sentiment will be any rise or descent for the true positive value of the negative in one interval with the “opposite” fluctuation of the Bitcoin price in the next shift. For example, if the true positive value of the negative class raised up and the Bitcoin price dropped in the next shift, this will be counting as an event for negative sentiment.

7 EXPERIMENTAL RESULTS

7.1 Sentiment Learning

Several single classifiers were trained and tested to evaluate their accuracies; **Table 4** shows the results of accuracies of each classifier on the trained data. The results show converge on classifiers results, except Naïve Bayes algorithm, which had the lowest results all datasets. On the other hand, decision tree methods perform better in most experiments, especially in N-Gram data models.

Modelled Dataset \ Classifiers		BN	NB	KNN	J48	RF
5Aug17	Tweet Embedding	50.98%	71.18%	50.34%	62.12%	66.43%
	N-Gram	57.06%	77.41%	62.74%	65.57%	68.04%
	Tweet Embs + N-Gram	52.25%	61.96%	52.44%	71.66%	66.22%
12Dec17	Tweet Embedding	76.13%	69.65%	54.31%	64.25%	85.92%
	N-Gram	85.37%	79.54%	50.88%	57.99%	89.33%
	Tweet Embs + N-Gram	82.28%	73.54%	51.02%	58.13%	88.97%
23Mar18	Tweet Embedding	81.92%	74.61%	49.74%	55.60%	88.79%
	N-Gram	86.56%	75.45%	70.03%	66.50%	88.93%
	Tweet Embs + N-Gram	85.51%	80.01%	71.62%	67.11%	89.58%
12Jul18	Tweet Embedding	52.45%	73.78%	48.39%	51.01%	69.07%
	N-Gram	66.40%	81.21%	41.32%	39.92%	73.53%
	Tweet Embs + N-Gram	67.14%	80.02%	42.54%	41.60%	71.81%
15Mar19	Tweet Embedding	77.83%	67.64%	54.25%	67.72%	86.56%
	N-Gram	82.48%	82.01%	66.92%	63.73%	88.75%
	Tweet Embs + N-Gram	85.12%	85.59%	58.19%	62.11%	89.12%

Table 4 The overall accuracies of each datasets by each machine learning classical classifier

For the deep learning classifiers, MLP and WiSARD shows a significant improvement in N-Gram, and tweet embedding with N-Gram data models, and it shows that manual sentiment performs better than RATS in tow datasets, but RATS gives more stable performance overall dataset. See **Table 5**.

Modelled Dataset \ Classifiers		MLP	RBFN	WiSARD
5Aug17	Tweet Embedding	62.78%	74.82%	58.15%
	N-Gram	64.91%	83.58%	61.95%
	Tweet Embs + N-Gram	66.49%	86.12%	62.55%
12Dec17	Tweet Embedding	84.39%	81.07%	73.96%
	N-Gram	91.26%	84.97%	76.25%
	Tweet Embs + N-Gram	91.033%	85.32%	77.64%
23Mar18	Tweet Embedding	88.40%	84.49%	58.76%
	N-Gram	91.82%	86.12%	87.29%
	Tweet Embs + N-Gram	92.97%	88.26%	87.45%
12Jul18	Tweet Embedding	68.56%	77.83%	59.43%
	N-Gram	70.97%	84.93%	62.73%
	Tweet Embs + N-Gram	71.15%	85.84%	60.96%
15Mar19	Tweet Embedding	82.94%	83.49%	68.49%
	N-Gram	81.37%	83.22%	82.73%
	Tweet Embs + N-Gram	83.97%	83.38%	70.35%

Table 5 The overall accuracies of the datasets by each deep learning algorithm

In general, all experiments are performed 5-folds cross validation test. The experiments show a higher degree of accuracy of deep learning algorithms over the classical machine learning classification methods. Manual sentiment gives better results in some datasets, while RATS is efficient in most experiments except with Naïve Bayes and Bayesian Network algorithms. Best accuracies were with manual sentiment using MLP deep learning algorithm and WiSARD, and with Tweet-Embedding with N-Gram data modeling. Therefore, the Tweet-Embedding with N-Gram modeling features will be applied in the time-series experiments.

To evaluate the results, several measures are considered beside the algorithm accuracy: precision, recall and F-1 score, where:

$$\text{Recall} = \frac{\sum(\text{True Positive})}{\sum(\text{False Negative}) + \sum(\text{True Positive})}$$

$$\text{Precision} = \frac{\sum(\text{True Positive})}{\sum(\text{False Positive}) + \sum(\text{True Positive})}$$

$$\text{F1-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

The results of these measures of all applied machine learning algorithms over all datasets, are shown in **Table 6** for classical machine learning algorithms, and **Table 7** for deep learning algorithms, focusing on positive and negative classes. The tables have a heat-map visualization feature with bold font for values that ranged from 80% to 100%.

Algorithm		BN			NB			KNN			J48			RF		
		Positive	Negative	Overall	Positive	Negative	Overall	Positive	Negative	Overall	Positive	Negative	Overall	Positive	Negative	Overall
5Aug17	Precision	0.871	0.201	0.827	0.456	0.174	0.746	0.740	0.677	0.853	0.622	0.703	0.853	0.966	1.000	0.882
	Recall	0.328	0.700	0.717	0.408	0.683	0.620	0.663	0.581	0.859	0.643	0.559	0.861	0.530	0.379	0.866
	F1	0.476	0.312	0.738	0.431	0.278	0.663	0.708	0.625	0.855	0.717	0.475	0.850	0.685	0.550	0.849
12Dec17	Precision	0.649	1.000	0.836	0.515	0.271	0.745	0.819	0.879	0.887	0.748	0.798	0.866	0.994	0.995	0.898
	Recall	0.646	0.455	0.823	0.620	0.858	0.518	0.735	0.802	0.889	0.695	0.823	0.869	0.566	0.683	0.881
	F1	0.648	0.626	0.811	0.563	0.412	0.545	0.774	0.839	0.887	0.721	0.792	0.867	0.721	0.810	0.872
23Mar18	Precision	0.923	0.985	0.871	0.868	0.374	0.781	0.934	0.923	0.899	0.926	0.849	0.906	0.917	0.999	0.912
	Recall	0.818	0.681	0.655	0.740	0.826	0.714	0.878	0.826	0.896	0.925	0.809	0.906	0.923	0.744	0.909
	F1	0.867	0.805	0.855	0.811	0.515	0.731	0.906	0.872	0.896	0.926	0.828	0.906	0.920	0.853	0.908
12Jul18	Precision	0.548	0.573	0.681	0.576	0.205	0.625	0.656	0.577	0.713	0.656	0.575	0.711	0.776	0.787	0.755
	Recall	0.645	0.430	0.677	0.354	0.749	0.425	0.623	0.519	0.718	0.627	0.502	0.716	0.558	0.438	0.749
	F1	0.593	0.491	0.677	0.439	0.322	0.458	0.639	0.547	0.715	0.641	0.536	0.713	0.649	0.563	0.734
15Mar19	Precision	0.884	0.796	0.852	0.515	0.366	0.639	0.885	0.841	0.891	0.865	0.830	0.884	0.954	0.996	0.910
	Recall	0.754	0.746	0.851	0.570	0.623	0.582	0.862	0.824	0.891	0.864	0.809	0.885	0.818	0.750	0.900
	F1	0.814	0.770	0.849	0.541	0.461	0.596	0.873	0.832	0.891	0.865	0.819	0.884	0.881	0.856	0.898

Table 6 shows the results of precision, recall and F-measure for each classical machine learning algorithms (lazy learner method, Bayesian methods, and decision tree methods). Datasets 12Dec17 and 23Mar18 (the short-time datasets) show better and more balanced precisions, recall and F-1 measures results, and dataset 15Mar19 shows the most balanced results among the 24-hours datasets, while Random Forest decision tree algorithm show the best scores on these evaluation measures.

Dataset	Algorithm	MLP			RBFN			WIS		
		Positive	Negative	Overall	Positive	Negative	Overall	Positive	Negative	Overall
5Aug17	Precision	0.572	0.703	0.853	0.640	1.000	0.751	0.606	0.965	0.829
	Recall	0.643	0.359	0.861	0.259	0.015	0.757	0.686	0.344	0.818
	F1	0.717	0.475	0.850	0.357	0.029	0.704	0.644	0.507	0.813
12Dec17	Precision	0.838	0.878	0.910	0.915	0.712	0.800	0.999	1.000	0.991
	Recall	0.817	0.875	0.910	0.106	0.826	0.781	0.968	0.973	0.991
	F1	0.828	0.877	0.910	0.190	0.764	0.735	0.984	0.987	0.991
23Mar18	Precision	0.946	0.911	0.930	0.824	0.737	0.838	0.998	1.000	0.990
	Recall	0.931	0.865	0.930	0.853	0.765	0.836	0.985	0.972	0.989
	F1	0.938	0.888	0.930	0.735	0.734	0.836	0.991	0.986	0.989
12Jul18	Precision	0.739	0.732	0.717	0.538	0.537	0.559	0.544	0.535	0.579
	Recall	0.477	0.343	0.712	0.331	0.099	0.610	0.167	0.109	0.600
	F1	0.579	0.467	0.688	0.410	0.167	0.558	0.255	0.181	0.520
15Mar19	Precision	0.825	0.945	0.846	0.718	0.481	0.708	0.593	0.804	0.805
	Recall	0.771	0.608	0.840	0.588	0.553	0.704	0.918	0.705	0.754
	F1	0.797	0.740	0.835	0.647	0.515	0.703	0.720	0.751	0.759

Table 7 show the results of precision, recall and F-measure for each deep learning algorithms. Datasets the short-time datasets (12Dec17 and 23Mar18) shows better, recall and F-1 measures results, and MLP with WiSARD show better scores on these evaluation measures.

6.2 Time Series Analysis

Relying on the true positive values of the positive sentiment class and the negative sentiment class, a collection of visualization charts has been built to visualize and note the events on the applied time-series. An event is the matching of rise and descent of the algorithm true positive value in an interval, with the rise and descent of the Bitcoin price on the next shift. Figure 2 and Figure 3 present a sample of what the time series charts look like. The figures show that fluctuations of the sentiment classes occur more events than the percentage of positive and negative tweets fluctuations between the intervals and compare its fluctuation with the Bitcoin price fluctuation. Bitcoin historical price data were found in bitcoincharts.com and it contain the BTC/USD rate for every minute. Using smoothing, BTC price in each interval was set up to closing price at the end of the interval. Moreover, single and double shifts are considered to find events in each dataset.

We can realize from the charts that there are some matchings in both positive and negative sentiment class accuracy and can be used to expect the price changing. For example, in Figure 2 (that present the fluctuation of true positive accuracy over the 12Dec17 intervals) in intervals number 4,5,6, and 7, positive sentiment accuracy had strong value of event rate because seven to eight of the used algorithms almost predict the right fluctuation, while in the negative sentiment accuracy, intervals number 4,6 and 7 doesn't had the same strong value. On the other hand, intervals number 10 and 11 in the negative sentiment accuracy had a very strong rate, but it has very weak value in positive sentiment accuracy for the same intervals. Therefore, considering both sentiments is useful to predict the fluctuation.

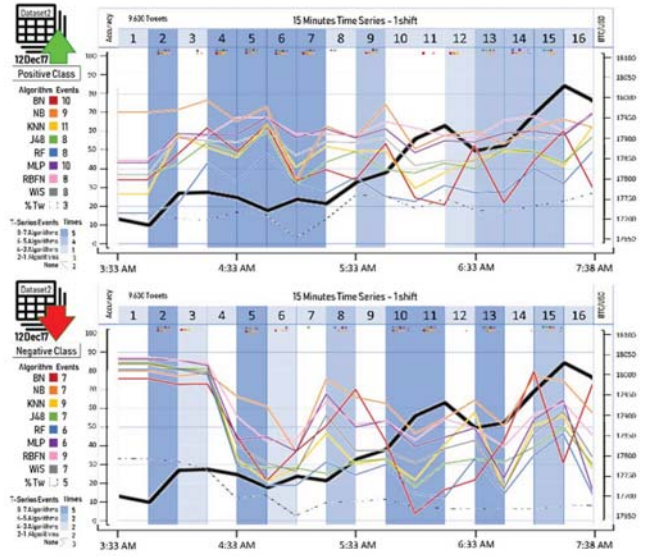


Figure 2 Time-series with 15 minutes interval chart of the second dataset and one shift to the future 15 minutes and compare it with the price fluctuations.

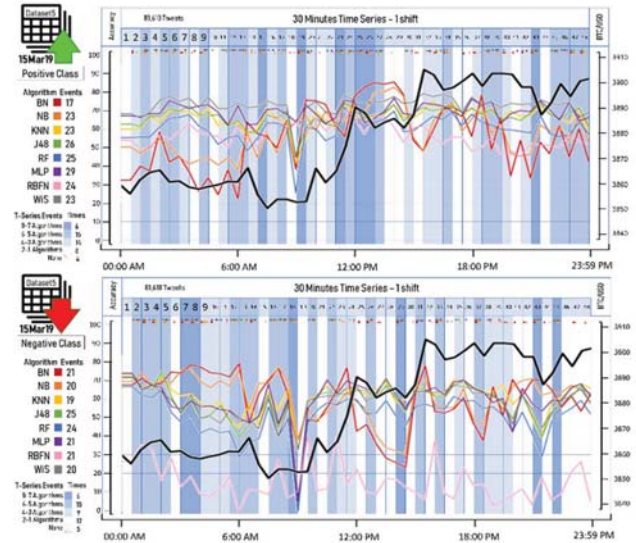


Figure 3 Time-series of 30 minutes interval chart of the fifth dataset one shift to the future 30 minutes, in comparison to price fluctuations.

See Table 8 and Table 9 to see overall events in all datasets with all shifts and intervals, and the percentages of all algorithms that have 50% or more of the predicted events. The results show a preponderance of the events of the positive class sentiment over the negative class sentiment in matched events. The results also showed that Naïve Bayes algorithm had a higher percentage of predicted events on both positive and negative sentiment together.

Algorithm			BN		NB		KNN		J48		RF		MLP		RBFN		WiS		Total	
			P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N
5Aug17	30MinTS	1Shift	24	25	25	23	19	22	24	20	26	19	26	21	22	19	20	20	186	169
		2Shifts	21	22	23	22	25	23	16	23	20	24	26	24	26	26	28	27	185	191
	60MinTS	1Shift	10	9	8	14	15	12	14	12	9	11	12	10	12	11	11	9	91	88
		2Shifts	11	12	10	10	8	12	7	11	5	12	9	11	10	9	7	12	67	89
	120MinTS	1Shift	5	4	5	3	7	4	7	3	7	4	8	6	10	3	7	3	56	30
		2Shifts	7	4	5	5	6	5	4	5	6	2	6	3	4	5	5	5	43	34
12Dec17	15MinTS	1Shift	10	7	9	7	11	9	8	7	8	6	10	6	8	9	8	7	72	58
		2Shifts	4	5	5	7	8	6	7	6	5	7	7	7	7	6	7	6	50	50
23Mar18	5MinTS	1Shift	7	9	11	10	12	13	8	8	8	10	10	11	10	8	10	8	76	77
		2Shifts	9	9	9	12	9	11	9	9	8	12	10	11	9	10	11	12	74	86
12Jul18	30MinTS	1Shift	26	24	22	18	24	25	24	26	22	21	22	23	22	28	26	190	185	
		2Shifts	24	16	22	24	21	27	24	26	25	28	21	22	23	23	25	183	191	
	60MinTS	1Shift	11	10	11	15	12	15	9	16	12	13	7	13	14	12	10	86	108	
		2Shifts	15	12	12	10	11	9	9	10	9	8	15	9	9	11	9	10	89	79
	120MinTS	1Shift	7	2	9	5	7	4	7	4	7	5	5	4	6	3	4	3	52	30
		2Shifts	4	6	2	6	5	4	3	4	5	5	7	6	4	5	8	7	38	43
15Mar19	30MinTS	1Shift	17	21	23	20	23	19	26	25	25	24	29	21	24	21	23	20	190	171
		2Shifts	26	23	29	27	17	26	20	26	23	20	28	24	25	24	25	183	203	
	60MinTS	1Shift	12	13	13	17	7	11	7	11	9	12	11	14	10	17	6	13	75	108
		2Shifts	9	10	12	12	15	9	15	7	18	10	11	10	16	6	16	7	112	71
	120MinTS	1Shift	6	4	5	3	5	4	6	3	6	4	7	6	7	3	6	3	48	30
		2Shifts	7	5	6	5	3	4	4	5	4	4	4	4	3	3	6	3	37	33
Total			272	252	276	275	270	274	258	267	267	264	283	270	280	257	277	265	2183	2124

Table 8 All events that predict the fluctuation for both positive and negative sentiment class accuracy using eight different machine learning algorithms

Dataset		Algorithm		BN		NB		KNN		J48		RF		MLP		RBFN		WIS		# Alg. >= 50%	
		P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N		
5Aug17	30MinTS	1Shift	52%	54%	54%	50%	41%	48%	52%	43%	57%	41%	57%	46%	48%	41%	43%	43%	5	2	
		2Shifts	44%	48%	50%	48%	54%	50%	35%	50%	43%	52%	57%	52%	57%	57%	61%	59%	5	5	
	60MinTS	1Shift	45%	41%	36%	64%	48%	55%	64%	55%	41%	50%	55%	45%	55%	50%	41%	5	5		
		2Shifts	50%	55%	45%	45%	36%	55%	32%	50%	23%	55%	41%	50%	45%	41%	32%	55%	1	5	
	120MinTS	1Shift	50%	40%	50%	30%	70%	40%	70%	30%	70%	40%	80%	60%	100%	30%	70%	30%	8	1	
		2Shifts	70%	40%	50%	50%	40%	50%	40%	50%	40%	50%	40%	60%	30%	40%	50%	50%	6	5	
12Dec17	15MinTS	1Shift	37%	50%	64%	50%	79%	64%	57%	50%	57%	43%	71%	43%	57%	64%	57%	50%	8	5	
		2Shifts	29%	36%	36%	50%	57%	43%	50%	43%	36%	50%	50%	50%	50%	43%	50%	43%	5	3	
23Mar18	5MinTS	1Shift	39%	50%	61%	56%	67%	72%	44%	44%	44%	56%	56%	61%	56%	44%	56%	44%	5	5	
		2Shifts	50%	50%	50%	67%	50%	61%	50%	50%	44%	67%	56%	61%	50%	56%	61%	67%	7	8	
12Jul18	30MinTS	1Shift	57%	52%	48%	39%	52%	54%	52%	57%	48%	46%	48%	50%	48%	48%	61%	57%	4	5	
		2Shifts	52%	33%	48%	52%	46%	59%	52%	57%	54%	61%	44%	48%	50%	50%	54%	5	6		
	60MinTS	1Shift	50%	45%	50%	68%	55%	68%	41%	73%	55%	59%	32%	59%	64%	55%	45%	64%	5	7	
		2Shifts	48%	55%	55%	45%	50%	41%	41%	45%	41%	34%	48%	41%	41%	50%	41%	45%	4	2	
	120MinTS	1Shift	39%	20%	90%	50%	70%	40%	70%	40%	70%	50%	50%	40%	60%	30%	40%	30%	7	2	
		2Shifts	40%	60%	20%	60%	50%	40%	30%	40%	50%	50%	70%	60%	40%	50%	80%	70%	4	6	
15Mar19	30MinTS	1Shift	37%	44%	50%	43%	50%	41%	57%	54%	54%	52%	63%	46%	52%	46%	50%	43%	7	2	
		2Shifts	57%	50%	63%	59%	37%	57%	43%	57%	50%	50%	43%	61%	52%	54%	52%	54%	5	7	
	60MinTS	1Shift	55%	59%	59%	77%	32%	50%	32%	50%	41%	55%	50%	64%	45%	77%	27%	59%	3	8	
		2Shifts	41%	45%	55%	55%	68%	41%	68%	32%	82%	45%	50%	45%	73%	27%	73%	32%	7	1	
	120MinTS	1Shift	40%	40%	50%	30%	50%	40%	40%	30%	60%	40%	70%	60%	70%	30%	60%	30%	8	1	
		2Shifts	70%	50%	60%	50%	30%	40%	40%	50%	40%	40%	30%	30%	30%	60%	30%	3	3		
Total of Events >= 50% per Algorithm				15	10	16	15	16	12	12	13	12	13	16	12	14	10	16	11	117	94

Table 9 Shows the percentages of all algorithms that have 50% or more of the predicted events.

To find correlation between the class sentiment accuracy and the Bitcoin price, two evaluation tests are applied: the student's t-test and the Chi-square test. **Table 10** and **Table 11** shows the associated p values of the positive class sentiment with the Bitcoin price for the machine learning classification methods and the deep learning methods. To confirm a statistical significance, a p value at the 0.05 probability level was determined. The results show a higher correlation in t-test using WiSARD deep learning algorithm than other algorithms. The tables also show that 5Aug17 with 30-minutes interval time-series and 23Mar18 with 5-minutes interval time-series had a signature p value with all algorithms in both t-students test and Chi-square test.

Dataset \ Algorithm		BN		NB		KNN		J48		RF	
		t test	Chi sq	t test	Chi sq	t test	Chi sq	t test	Chi sq	t test	Chi sq
5Aug17	30MinTS	8.62e-6	0.0002	0.0355	3.3e-5	0.0061	1.85e-8	0.0022	1.03e-8	1.74e-5	1.76e-7
	60MinTS	0.0036	0.0329	0.0649	0.0055	0.2105	0.0055	0.1342	0.0009	0.0718	0.0014
	120MinTS	0.0518	0.0947	0.2169	0.0703	0.3547	0.0777	0.3272	0.0797	0.0556	0.0169
12Dec17	15MinTS	0.3921	0.0011	0.0003	0.2315	0.1086	0.0003	0.1359	0.0003	0.1236	1e-6
23Mar18	5MinTS	0.0002	0.0001	0.0009	2e-6	0.2375	8.25e-7	0.0069	3.5e-5	2.55e-6	0.0007
12Jul18	30MinTS	0.0195	0.9499	0.1194	0.0014	0.0009	0.0069	0.0042	0.0036	0.1497	2.4e-5
	60MinTS	0.0663	0.0772	0.3732	3.9e-5	0.0274	0.0396	0.0295	0.0449	0.2086	0.0067
	120MinTS	0.0795	0.3026	0.4528	0.0151	0.0828	0.2535	0.1495	0.3156	0.2470	0.1531
15Mar19	30MinTS	0.1957	0.9522	0.0693	0.9809	0.0013	0.9999	0.0009	0.9999	0.0274	0.9953
	60MinTS	0.2308	0.2975	0.1721	0.1803	0.0037	0.8823	0.0035	0.9205	0.0229	0.7013
	120MinTS	0.2782	0.1749	0.4144	0.0586	0.0275	0.8026	0.0200	0.8384	0.0649	0.6639

Table 10 The p values result of t-test and Chi-square test of the positive sentiment accuracy with the Bitcoin price for the machine learning classification methods.

Dataset \ Algorithm		MLP		RBFN		WiS	
		t.test	Chi sq.	t.test	Chi sq.	t.test	Chi sq.
5Aug17	30MinTS	0.0298	2e-6	0.0154	1.74e-7	0.0094	2e-6
	60MinTS	0.4081	0.0158	0.0811	0.0080	0.2363	0.0079
	120MinTS	0.4779	0.1074	0.1884	0.0815	0.3357	0.0647
12Dec17	15MinTS	0.0018	0.0128	0.0006	0.0209	0.0141	0.0017
23Mar18	5MinTS	8.47e-6	0.0002	3.99e-5	6.2e-5	4.78e-7	0.0007
12Jul18	30MinTS	0.0125	0.0019	0.3669	2.14e-7	0.0003	0.4369
	60MinTS	0.1545	0.0269	0.2626	0.0006	0.0354	0.0682
	120MinTS	0.2644	0.1455	0.2773	0.0363	0.1856	0.2329
15Mar19	30MinTS	0.0002	0.9999	0.1047	0.9979	1.85e-5	1
	60MinTS	0.0049	0.8952	0.3124	0.4291	4.56e-4	0.9850
	120MinTS	0.0808	0.6617	0.3886	0.1804	0.0160	0.8961

Table 11 The p values result of t-test and Chi-square test of the positive sentiment accuracy with the Bitcoin price for the deep learning methods.

The evaluation tests showed a partial correlation between the fluctuation of the sentiment classes using different machine learning algorithms and the fluctuation of the Bitcoin price. Short intervals charts have more correlated values to the price fluctuation. The scores confirm that MLP, WiSARD, and decision tree methods have better correlation among all used algorithms in the study.

7 CONCLUSION

Tweet sentiment analysis is an active field for studies in price forecasting. Using Twitter in sentiment analysis for Bitcoin is becoming an important step for most researchers due to the large amount of news feeds per minutes regarding Bitcoin. Therefore, text mining and classification techniques on Twitter data that can predict the best sentiment are needed. Classifying different datasets model using tweet-embeddings and N-Gram is useful to enhance prediction.

The main contribution of this paper is finding a partial correlation between the Bitcoin price fluctuation and the fluctuation of the sentiment classes using different machine learning algorithms, and providing a framework of an efficient tweet sentiment tool for Bitcoin tweets whether they are positive or not. In addition, the paper compares

different classification methods by providing detailed experimental evaluation. The results showed better accuracy for manual sentiment in the data preprocessing phase in some datasets, but it shows also a more stable performance overall experiment using R-Auto Tweets Sentiment (RATS). The results also showed that tweet-embedding and N-Gram data modeling features can improve the sentiment prediction, especially in MLP and WiSARD deep learning techniques, and decision trees algorithms. It is notable that having more features in the modeled data can improve the sentiment accuracy. This could be one of the challenges, especially with the large datasets, because some algorithms such as MLP deep learning or decision tree methods are time consuming experiments.

As a future study, designing a special lexicon for Bitcoin sentiment may improving the correlation of the sentiment analysis with the Bitcoin price fluctuation, with considering other features like hashtags, Twitter user, number of tweets, and emoticons.

REFERENCES

- [1] Hileman, G. and Rauchs, M. (2017). GLOBAL CRYPTOCURRENCY BENCHMARKING STUDY. Cambridge, United Kingdom: Cambridge Centre for Alternative Finance.
- [2] Pantas and Ting (2017). Blockchain Technology, Beyond Bitcoin. Berkeley, California: Berkeley Engineering, Sutardja Center for Entrepreneurship & Technology Technical Report.
- [3] Kaggle. (2019). Datasets | Kaggle. Available at: <https://www.kaggle.com/datasets>. [Accessed: 13- October- 2019].
- [4] Abirami, R. and Vijaya, M.S.. (2012). Stock Price Prediction Using Support Vector Regression. Communications in Computer and Information Science Journal. 588-597
- [5] G. Holmes; A. Donkin; I.H. Witten (1994). "Weka: A machine learning workbench" (PDF). Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.
- [6] V. Pagolu, K. Challa, G. Panda and B. Majhi, Sentiment Analysis of Twitter Data for Predicting Stock Market Movements, International conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), 2016.
- [7] Matta, M. and Lunesu, I. Marchesi, M. Bitcoin Spread Prediction Using Social and Web Search Media. Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management – 2015
- [8] Stenqvist, E. and Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis.
- [9] Sagar, C. (2018). Twitter Sentiment analysis using R. [online] Dataaspirant. Available at: <http://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/> [Accessed 13 Oct. 2019].
- [10] Krafft, P., Penna, N and Pentland, A. (2018) An Experimental Study of Cryptocurrency Market Dynamics. ACM CHI Conference on Human Factors in Computing Systems (CHI)
- [11] L. Torgo, Data Mining with R, learning with case studies, Boca Raton, FL, Champan & hall/CRC, 2010.
- [12] N. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression ". The American Statistician, vol. 46, no. 3, p. 175, 1992.
- [13] Ruggeri F., Faltin F. & Kenett R., Bayesian Networks - Encyclopedia of Statistics in Quality and Reliability. (2008).
- [14] S. Russell and P. Norvig, "A modern, agent-oriented approach to introductory artificial intelligence", ACM SIGART Bulletin, vol. 6, no. 2, pp. 24-26, 1995.
- [15] Salzberg, S. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, 16(3), pp.235-240.
- [16] L. Breiman, "Random Forests". Machine Learning 45 (1): 5–32. 2001.
- [17] Witten, I. (2019). More Data Mining with Weka - Simple neural networks. New Zealand: Department of Computer Science University of Waikato.
- [18] Frank, E. (2014). Fully Supervised Training of Gaussian Radial Basis Function Networks in WEKA. Department of Computer, Science University of Waikato.
- [19] De Gregorio, M. and Giordano, M. (2017). The WiSARD Classifier. ESANN 2016.