National College of Ireland

Prediction of Cryptocurrency Price based on Sentiment Analysis and Machine Learning Approach

MSc Research Project
Data Analytics

Sai Prasanna Gontyala
Student ID: X19233388

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sai Prasanna gontyala
Student ID:	X19233388
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/08/2021
Project Title:	Prediction of Cryptocurrency Price based on Sentiment Analysis and Machine Learning Approach
Word Count:	6037
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Cryptocurrency Price based on Sentiment Analysis and Machine Learning Approach

Sai Prasanna gontyala
X19233388

Abstract

Due to the significant expansion in the field of social media, sentiment analysis is playing an increasingly important role in the technical world. Sentiment analysis is motivated by the fact that social media platforms such as Twitter provide a wonderful forum for the general public to voice their opinions about a product or an event. Such viewpoints allow academics to work on data mining based on public reviews and opinions, and they provide crucial insights that aid corporations in making better decisions.Also, cryptocurrencies have grown in popularity in recent years and are now accepted in nearly all nations. The derivation of emotion score, which specifies the intensity of each tweet containing the words "bitcoin" and "BTC," is discussed in this study. The sentiment analysis is performed on the twitter data using the python module vaderSentiment. Using the cross-correlation statistical techniques Spearman Correlation, Pearson Correlation, and Kendall Correlation, the relationship between the derived sentiment scores and the bitcoin time series data was determined.The LSTM model with two optimizers, Rmsprop and Adam, is used to predict the closing price of bitcoin. This would assist investors in understanding current public views around the world, and the bitcoin price would be affected as a result.

1 Introduction

Cryptocurrencies have gained tremendous popularity in recent years as a result of their high returns and proven potential. Transparency, traceability, low transaction costs, and big rewards on short-term investments have led to a surge in cryptocurrency investments. However, concerns about crypto traders' privacy and the cryptocurrency's high voltaic nature impact investors. Thanks to the advancements in data analytics, organizations have incorporated forecasts in every part of company function based on the user opinions recorded on various platforms, including social media.

Many elements influence the voltaic nature of cryptocurrency. User and public opinion also influence the price of cryptocurrencies. Users' satisfaction, happiness, or unhappiness can stimulate or demotivate their continued use of cryptocurrency. The prediction of crypto prices based on the opinion can give crypto users a sense of security, encouraging them to invest or transact more. The sentiments of cryptocurrency users can have an impact on bitcoin price changes. The right identification of sentiment scores would provide a solid framework for determining a virtually accurate bitcoin price, which is the subject of this study proposal. The goal of this study is to see how sentiment analysis may be used to anticipate the direction and accuracy of Bitcoin prices. A major study

of cryptocurrency research in Data Analytics field was conducted using time series data. There was limited research done previously which could correlate the public opinions (sentiments) with the price of the bitcoin and the influence on it.

Since the evolution of the first crypto-coin, Bitcoin, cryptocurrency has had a significant impact on the financial world. It's been over a decade since there were more than 1700 cryptocurrencies available to the public, two of which are traded on major platforms. Cryptocurrencies are becoming increasingly popular because to their numerous benefits, which include being fast, safe, scalable, trustworthy, reliable, decentralized, and more. The desire to hold this digital cash has turned it into a commodity rather than a currency. The profits are substantial, as they have been for cryptocurrency investors in the past few years. 2017 was the year when Bitcoin resurrected from the ashes, reaching a high of nearly 20,000 dollars at the end of the year paving the way for new billionaires to make their fortunes by investing in the cryptocurrency. This resulted in a frenzy among investors as well as the general public. This was the beginning of the bitcoin investing craze, as everyone wanted to put their money on the line. There are various types of cryptocurrencies available today, each with its own set of technologies and features. Due to its appeal among investors, Bitcoin continues to lead the market capital charts today. Alternate currencies were designed as an alternative to blockchain technology, in which, in addition to incorporating the genetics of being a cryptocurrency, they also incorporate the genetics of being a digital asset. Majority of them were introduced to address the practical use cases. While bitcoin was introduced mainly for payments. Furthermore, projects like Taipei's smart city and Japan's energy trading network, which uses the Iota cryptocurrency, provide investors more traction and hope for bigger profits.

1.1 Research Question

Can the price of the cryptocurrency be predicted based on the sentiment scores determined from the twitter data that is merged with the historical time series price data to enhance the field of cryptocurrency and support investors in making better investment decisions?

1.2 Research Objectives and Contributions

The research project objectives below are proposed with respect to the research questions posed in this project with a set road map for completing this project successfully.

Objective 1: The first objective is to source the price data for Bitcoin and extract the tweets related to "bitcoin" and "BTC" using twitter API.

Objective 2: Cleaning the tweets and removal of the special characters and unwanted characters as a part of pre-processing.

Objective 3: Determining the sentiment score for the tweets extracted using Text Blob in python.

Objective 4: Using the statistical correlational techniques Spearman, Pearson and Kendall to plot the correlation of sentiment score and closing price of bitcoin.

Objective 5: At this stage the algorithms are implemented for predictive modelling and evaluating their results.

Objective 6: This is the final stage where the prediction results of different algorithms would be compared for different cryptocurrencies, fulfilling our main research question.

The main objective for performing the above steps is to predict the bitcoin price and to identify the relation between bitcoin price and sentiment scores of the twitter data.

Contributions:

The gaps identified in the literature assessment and the importance of the research topics posed are two ways in which this research project contributes to the body of knowledge.

Below are the contributions of the project based on the objectives.

- Sentiment score calculation for each tweet related to "bitcoin" and "BTC" using Vader Sentiment approach.
- Correlation analysis for calculated sentiment score and price of bitcoin using the statistical techniques - Spearman Correlation, Pearson Correlation and Kendall Correlation.
- Prediction of bitcoin price based on the sentiments derived from the tweets using LSTM model.

The next sections of the report are illustrated as following. Chapter 2 presents the related work of cryptocurrencies price prediction using Machine Learning, Sentiment Analysis using textual data and Text Pre-processing that's usually employed during Data Mining and Sentiment Analysis. Chapter 3 illustrates the CRISP-DM methodology used in the implementation of the project. Chapter 4 presents the design and architecture of the project. Chapter 5 presents the step-by-step implementation of the project that covers data preparation, data pre-processing, sentiment score calculation, correlation analysis and predictive model build. Chapter 6 describes the evaluation of the model and the visualizations. Chapter 7 presents the conclusion of the implemented work and the future scope of it.

2 Related Work

Many researchers have been paying attention to the rise in the volume of trade and the value of cryptocurrencies in recent years. There have been several attempts to anticipate the price of bitcoin using various indicators such as transaction volume and previous price data.

2.1 Machine Learning in Cryptocurrency

Cryptocurrency price predictions is now in its early stages. The majority of the study done so far has focused on short-term price prediction based on various factor. Artificial Neural Network (ANN) and Long-Short-term-Memory (LSTM) neural networks was used to estimate the price of the cryptocurrencies Bitcoin, Ethereum, and Ripple using historical price data from 7th August 2015 to 2nd June 2018 time series data (Yiying and Yeze; 2019). The results show that the ANN excelled at long-term prediction while the LSTM excelled at short-term prediction. They did say, however, that for the analysis, 30-day historical data would have been more informative. We are examining dataset of time series for a timeframe of one month for the current method.

The factors that drive cryptocurrencies utilizing a variety of trading tactics, including fundamental, multifactor, trend monitoring, and arbitrage strategies were highlighted

(Sun et al.; 2019). The data that is used in the analysis is of time-series at a frequency of 5 minutes from August 2017 to December 2018. To extract features from the 5-minute frequency time-series data, they used 16 variables in Alpha101 for feature engineering. The validity of the test of variables, however, is not addressed in their study. Instead, they chose random forest for classification, which selects components at random and then integrates the findings to get classification results. The model attained good results when the prediction was for a long term than for a shorter time interval. However, due to the voltaic nature of crypto trading, accurate findings over a short time frame are critical.

Martin et al. (2020) forecasted bitcoin price using machine learning techniques. For the analysis, time series data was utilized. Following a principal component analysis (PCA), it was decided to employ the components Market cap (daily volume multiplied by daily price), volume, and capitalization change 1 day for further investigation. The ensemble technique under the weighted system produced the best accurate predictions out of the numerous approaches they used. The analysis was entirely based on historical data.

Volatile nature of the cryptocurrencies is one of the major concerns for the investors. In the study of the volatile nature of cryptocurrencies - Bitcoin, Ethereum, and XRP, the time series data was used (Saadah and Whafa; 2020) . To check the stability of these cryptocurrencies, the researchers devised three algorithms: K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). The execution time of each algorithm, as well as its accuracy and RMSE, were focussed in the study. The LSTM approach outperformed the other algorithms in terms of accuracy, although it took longer to execute.

2.2 Text Representation and Pre-processing

Text Representation is a crucial step in text mining, in which the data in documents is converted into numeric vectors that correspond to the texts. The performance of three document indexing methods: Term Frequency Inverse Document Frequency (TF-IDF), Latent Semantic Indexing, and Term Frequency Inverse Document Frequency (TF-IDF) (LSI) was compared (Zhang et al.; 2008). Information retrieval and text categorization are used to evaluate the performance. In comparison, the TF-IDF and multi-word algorithms performed well. A simple and intuitive linguistic strategy is used to extract many words. TF-IDF, on the other hand, works in two stages. It begins by determining the supremacy of terms in a particular document. The importance of the ranking term in the collection is then calculated.

Word distributed sensitive topic representation model (WDS-LDA), a topic modelling approach that uses human cognitive abilities and cognitive models to express topics was illustrated (Han et al.; 2021). This was accomplished by combining a weighted method with LDA. This improved the accuracy, but the computation speed and cost were extremely high, and because we're looking at a dataset containing a month's worth of data, the posts would be enormous, so this method was not considered for the analysis.

The influence of news stories on the bitcoin price, a novel text representation approach called SentiGraph was introduced (Yao et al.; 2019). Unlike N-Gram and Term Frequency Inverse Document Frequency (TF-IDF), the SentiGraph turns news articles' texts into a sentiment graph. This examines how one person's thoughts on a social network impact other, either positively or adversely. The colour of the nodes in the sentigraph obtained indicates the sentiment score. However TF-IDF has shown good performance in con-

verting texts to vectors based on the superiority of the terms.This would be useful when using twitter data for analysis, as it would allow to filter the retweeted data to see how individual opinions as a group affect the price.

2.3 Sentiment Analysis on Cryptocurrency

Emotions appear to play a significant impact in forming decisions (Ahn and Kim; 2020) and (Narman and Uulu; 2020). The sentiment analysis based on the comments posted by various users on Reddit related to cryptocurrencies was performed . They discovered that there is a relation between the price of cryptocurrencies and the quantity of comments. They also discovered that there is a relation between the mood of the commentators (both optimism and negativity) and the price change.

Knowing the feasibility information regarding areas for bitcoin investments around the world would be beneficial to investors (Bibi et al.; 2019). They offered an approach for developing a recommendation system that would assist investors in determining people's positive, negative, and neutral attitudes based on the places that have been identified as having a high adoption of bitcoin. The information is gathered from Twitter and then cleaned up. To get stem words and reduce text noise, Poster's algorithm was used to apply word stemming. The hashtags and subject modelling were then used to determine the locations. The data was subjected to sentiment analysis to determine the positive, negative, and neutral perceptions of the selected locations. For several cryptocurrencies, visualizations of people's sentiments based on location were captured in the study.

The change in public opinion and the volume of opinions made on a specific topic on the social media platforms can cause a direct or in-direct impact on the product. Balfagih and Keselj (2019) examined the relationship between sentiment analysis on Twitter data about bitcoin and price fluctuations. They improved classification accuracy by combining Twitter embedding and N-Gram data modelling techniques. They had also tested the accuracy of two Twitter sentiment approaches: manual sentiment and auto-tweet sentiment. Tweets were manually classified as positive, negative, or neutral in the manual sentiment approach. The getnrcsentiment function in R is used to assign a sentiment score to each tweet in the auto-tweet sentiment technique. The results conveyed that the best accuracies were attained using Twitter embedding and N-Gram data modelling approaches on decision tree algorithms.

The emotions of the people that had shown interest in the bitcoin discussions from the posts on bitcointalk.org using canonical sentiment analysis was quantified (Ahn and Kim; 2020). They had used Linguistic Inquiry and Word Count (LIWC) software to calculate percentage value for each post which is further used in empirical analysis to derive the various emotions that the posts conveyed. They had presented the statistics of the various emotions derived. However, the evaluation of the statistics was not quite justified using evaluation metrics.

The improvised version of TF-IDF weighting schemes to that of the standard TF-IDF was studied and the performance of the weighting scheme was evaluated (Al-Ghuribi et al.; 2020). On short texts, the updated version performed better. A change is made to the computation of Word Frequency (TF), which was determined by dividing the total number of sentences in the text by the number of times the term appears in the text information.

Document-Term Matrix (DTM) and Latent Dirichlet Allocation (LDA) for the prediction of Phishing Detection was used (Gualberto et al.; 2020). The LDA approach had

worked well and resolved the issues of high dimensionality as it extracts some discriminative features from the pre-processed texts.

3 Methodology

A modified version of CRISP-DM process model was followed to implement the current research. It is a six-step method for organizing data science research or machine learning projects. It is idealized as a series of events, although many actions can be accomplished in any order and with any number of iterations, depending on the developer's needs. In the process, the stages listed below will be followed.

Figure 1 helps in understanding the methodology followed in research project.

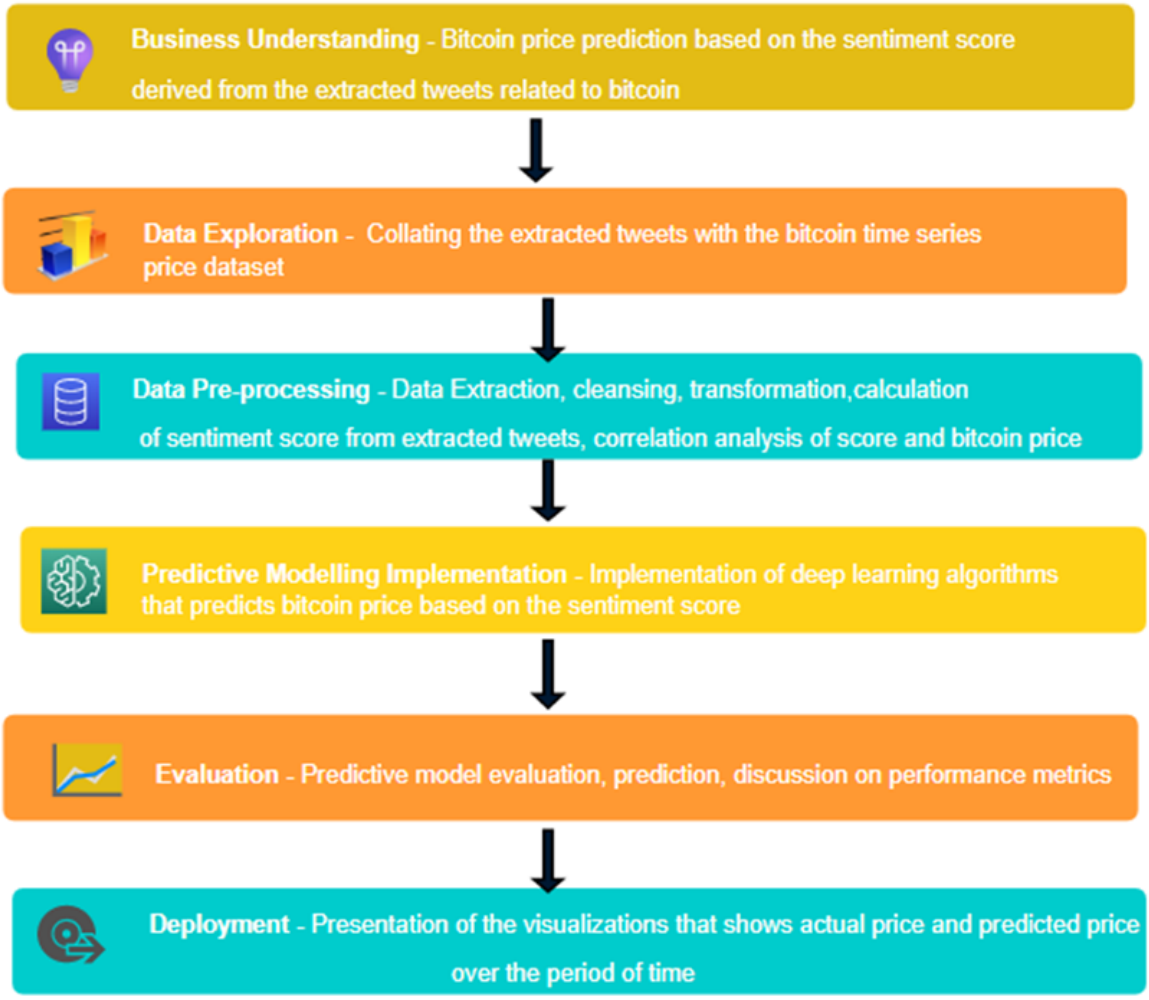

Figure 1: Bitcoin price prediction planning methodology

Business Understanding – Understanding of the Bitcoin price prediction based on the sentiment analysis of the tweets related to it and the machine learning approach.

Data Exploration – Exploration of data extracted from twitter and Correlation Analysis of bitcoin price data.

Data Pre-processing – Pre-processing of twitter data and determination of sentiment score from extracted tweets.

Predictive Modelling – Machine Learning model analysis for prediction of bitcoin price.

Implementation and Evaluation – Predictive model implementation and evaluation of the models.

Results Visualization – Visualization of the evaluated results and comparison of prediction and actual price of bitcoin.

4 Design Specification

The three-tier architecture was used to design the current research. Its applications are organized in three tiers – Presentation tier, Application tier and Data tier. The design for the current research is detailed below.

Tier 1 (Presentation) – The user interacts with the application through the presentation layer. It allows the user to interact with bitcoin forecasted price values in order to develop model output and show bitcoin price prediction insights.

Tier 2 (Application) – In the application layer, the sentiment score of each pre-processed tweet was calculated using VADER algorithm and aggregated based on the time-stamp. Correlation of the calculated sentiment score is checked using the statistical cross correlation techniques – Pearson, Kendall and Spearman. The predictive model LSTM is applied to predict the closing price of the bitcoin based on the sentiment score derived.

Tier 3 (Data) – The extraction of data is done in the data tier using the twitter API. Tweets from 11th May 2018: 9.00 AM to 29th May 2018: 12.00 PM were extracted using the python library. The corresponding bitcoin price for each hour is determined using the API crypto compare.

Figure 2 helps in understanding the design followed in research project.

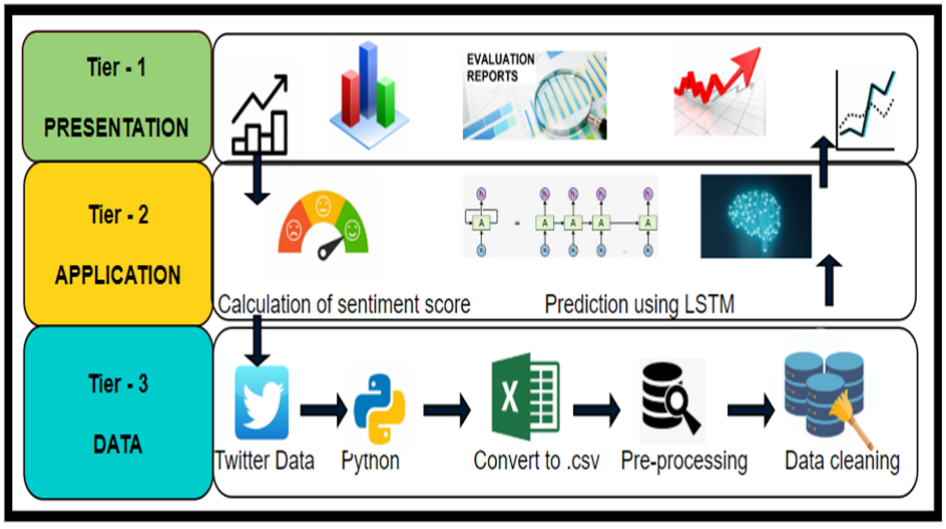

Figure 2: Design Specifications

5 Implementation of Cryptocurrency Prediction Models

This section illustrates the step-by-step implementation of the project. It describes the entire process of data preparation, data pre-processing, sentiment score calculation, correlation analysis and implementation of the predictive deep learning model to predict the bitcoin price from the sentiment score calculated.

5.1 Data Extraction

Data to implement the bitcoin price prediction project was extracted from two open sources.

5.1.1 Twitter

Twitter is a popular social media site where millions of people express themselves through "tweets." Likes and re-posts are used to convey support or disapproval of the tweets. The data was extracted using the Twitter API and the key terms "bitcoin" and "BTC" to study and comprehend people's attitudes towards bitcoin. To use python libraries to connect with the Twitter API, a premium developer account was first obtained to get a high access rate and access to old tweets. Tweets were collected from May 11th, 2018 at 9:00 a.m. to May 29th, 2018 at 12:00 p.m., using the "twython" library, which was created for extracting tweets using the Twitter API and authentication keys – Access token and Access token secret – that were provided.Id, Text, Username, Number of Followers, Number of Tweets, Number of Likes, and Tweet creation date and time details were collected for each tweet. Twitter authenticates with python to access its API, using the function "Twython" from the python package "twython" and oauthversion as 2.After successfully authenticating with the Access token and Access token secret provided by Twitter, the tweets can be retrieved using the streaming API or by specifying time frames for previous tweets. The extracted data is then translated to CSV format and used in the data pre-processing.

5.1.2 Bitcoin time-series data

Bitcoin is the world's first decentralized digital currency, as it operates without the use of a central bank or a single administrator. The system was created to function as a peer-to-peer network, in which transactions are made directly between users without the use of an intermediary. The bitcoin historic price data is extracted from an API, crytocompare [1]. Anyone can access and utilize the bitcoin history data for free. The hourly bitcoin price history for the corresponding time frame of tweets is extracted using the API. Closing price, High price, Low price, Opening price, and time stamp are the attributes of the dataset. The current implementation is a thought to be able to anticipate the hourly closing price.

[1] `https://min-api.cryptocompare.com/`

5.2 Data Pre-processing

The raw Twitter collection contains various undesired characters, photos, videos, and hashtags. During the pre-processing stage, these are eliminated as it impacts the sentiment score calculation against each tweet.

Figure 3 illustrates the removal of URLs, hashtags and special characters for tweets.

```
text = text.replace("#", "")
text = re.sub(
    r"https?://(?:[-\w.]|(?:%[\da-fA-F]{2}))+", "", text,
    flags=re.MULTILINE)
text = re.sub(r"@\w+ *", "", text, flags=re.MULTILINE)
```

Figure 3: Pre-processing of twitter data

- **Removal of URLs:** Users tend to include hyperlinks in their posts. The extracted tweets had URLs which would contribute the least to the calculation of sentiment score. In addition, more redundant data would hamper the computational speed and accuracy of the analysis.

- **Removal of hashtags:** Users tend to use hashtags to express their opinion in short or to indicate the tweet is related to a specific topic.Although the tweets with "BTC" are extracted, the character "." would not contribute anymore post the preparation of datasets for the calculation of sentiment score. So the hash character is removed from the extracted tweets data.

- **Removal of usernames:** Users can tag other users in the tweets. Usually this is used when a user intend to convey it about an other user but not a necessary though. As this would not contribute to the sentiments of the tweet, the usernames in the text of the tweet are identified using the character "@" and are removed from the text.

- **Removal of special characters:** Computational speed and performance of the analysis gets impacted as the volume of data increases. This is handled by removing the redundant data which do not contribute to the analysis. As the special characters like question mark ("?"), exclamation mark ("!"), semi colon (";") and "@" only increase the volume of the data for analysis and do not exhibit any polarity, such characters are removed in pre-processing.

- **Removal of white spaces:** Many Twitter users leave unneeded white spaces, which were deleted throughout the cleaning process.

- **Removal of duplicates:** Duplicates in the twitter data and bitcoin price data are checked. In the event that duplicates are found in either dataset, the duplicates are removed.

5.3 Sentiment score calculation

After the data is cleaned, pre-processed Twitter data and bitcoin data are loaded into two separate .csv files. The sentiment intensity of each text is calculated using VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm. The VADER algorithm is a combined approach of lexicon and rule-based sentiment analytic software. VADER is feasible to identity the polarity of text into three categories - positive, negative or neutral. It can also be used to calculate the intensity of sentiment of a text. It uses factors like emojis, intensifiers, contraction, punctuation and acronyms to calculate the scores. The positive value of the sentiment score indicates positive emotion in the tweet and the value of the score indicates the intensity of it. Negative value of the sentiment score indicates negative opinion and the value indicates the respective tweet intensity. While zero indicates the neutral sentiment.

The textual data of each preprocessed tweet is used to determine a sentiment score using VADER algorithm. Python provides a library "vaderSentiment" which is used to determine the score. Using "polarityscores" function in "SentimentIntensityAnalyzer" which is a part of "vaderSentiment" is used to determine the sentiment score against each tweet called compound. Figure 4 helps in understanding the calculation of compound score against each tweet using python libraries.

```
analyzer = SentimentIntensityAnalyzer()
compound = []
for i, s in enumerate(df_clean["Text"]):
    vs = analyzer.polarity_scores(s)
    compound.append(vs["compound"])
```

Figure 4: Compound score calculation using python libraries

As the analysis is aimed to know the impact of the sentiments of the people opinion on the price of the bitcoin and the further impact of a tweet can be calculated based on the followers and likes of the tweets, these factors are considered during the calculation of the final sentiment score. The score determined is equivalent to the compound score of the tweet multiplied by the user followers and the likes attained by the tweet. A famous personality or an influencer are most likely to have more followers and can get more like in comparison to an ordinary person which is a practical scenario. This is considered during the calculation of the sentiment score against each tweet. Figure 5 helps in understanding the formula used for the calculation of sentiment score using compound score determined, user followers count and likes of the tweet.

The calculated sentiment score is then grouped based on the time stamp of the tweet. This data is then converted into a csv file which has the attributes time and sentiment score that was calculated against each tweet. Closing price of the hour is chosen to b the response variable for the prediction of bitcoin price. All other attributes in the bitcoin price dataset High price, Low Price and Opening price are dropped. As the price prediction is designed to be made on an hourly basis, the data in both the datasets

```
    scores.append(s["compound"] *
                  ((s["UserFollowerCount"] + 1)) * ((s["Likes"] + 1)))
df_clean["score"] = scores
```

Figure 5: Sentiment score calculation formula

related to bitcoin and twitter are aggregated based on hour. These two dataframes are then merged using the time which is common attribute in both datasets. A final pre-processed dataframe with the attributes time, sentiment score and bitcoin price in USD is created for the prediction of bitcoin price for the next hour.

Figure 6 helps in understanding the formula used for the calculation of sentiment score using compound score determined, user followers count and likes of the tweet.

```
    scores.append(s["compound"] *
                  ((s["UserFollowerCount"] + 1)) * ((s["Likes"] + 1)))
df_clean["score"] = scores
```

Figure 6: Sentiment score calculation formula

Figure 7 helps in understanding the formula used for the calculation of sentiment score using compound score determined, user followers count and likes of the tweet.

```
    scores.append(s["compound"] *
                  ((s["UserFollowerCount"] + 1)) * ((s["Likes"] + 1)))
df_clean["score"] = scores
```

Figure 7: Sentiment score calculation formula

5.4 Correlation Analysis

The correlation analysis of the sentiment score determined from the tweets with the bitcoin price plays an important role in the prediction. The sentiment score should be correlated to the bitcoin price to be used as a variable in determining the price. It quantifies the relationship strength associated the derived sentiment score and the bitcoin hourly closing price. But in case of the analysis to find the impact on the sentiments of the people on the bitcoin price a mere correlation would be insufficient. As the rise or fall of the price can make an impact on the public opinion after the event. Also, the change in opinion of the public can later have an impact on the price. This marks the importance of the cross correlation analysis

The distinction is that cross-correlation introduces a lag, allowing one of the timeseries to be shifted left or right to obtain a better correlation. This fits with our purpose because the currency fluctuations happen after the sentiments in the tweets. As a result, we have made an analysis using the cross correlation. Three statistical correlation methods - Spearman, Pearson and Kendall were used and compared in the analysis.

Figure 8 shows the correlation between the derived sentiment score from the tweets and the bitcoin price.

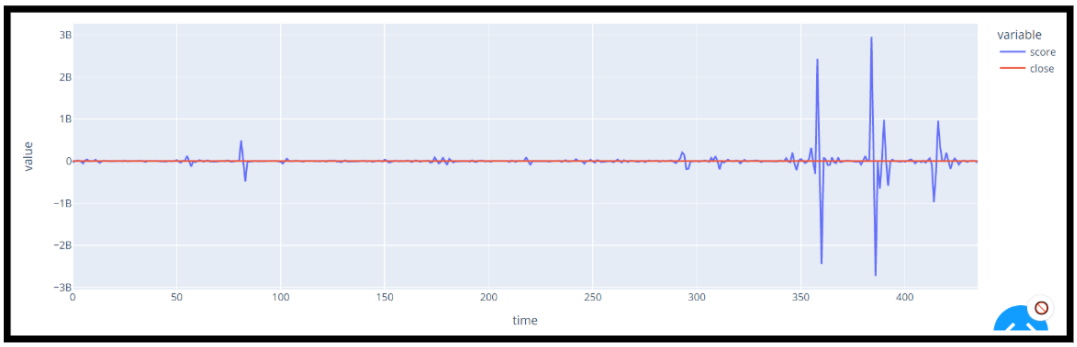

Figure 8: Derivative score and bitcoin price correlation

5.4.1 Spearman Correlation

The Spearman Rank Correlation is a non-parametric statistical technique for determining the degree of correlation between variables. It makes no assumptions about data distribution and only requires that the data be ordinal. The relation is seen to be positive with the time lag. The highest correlation is observed to be 0.076 when the lag is 11.

Figure 9 shows the cross correlation between the derived sentiment score from the tweets and the bitcoin price using Spearman correlation method.

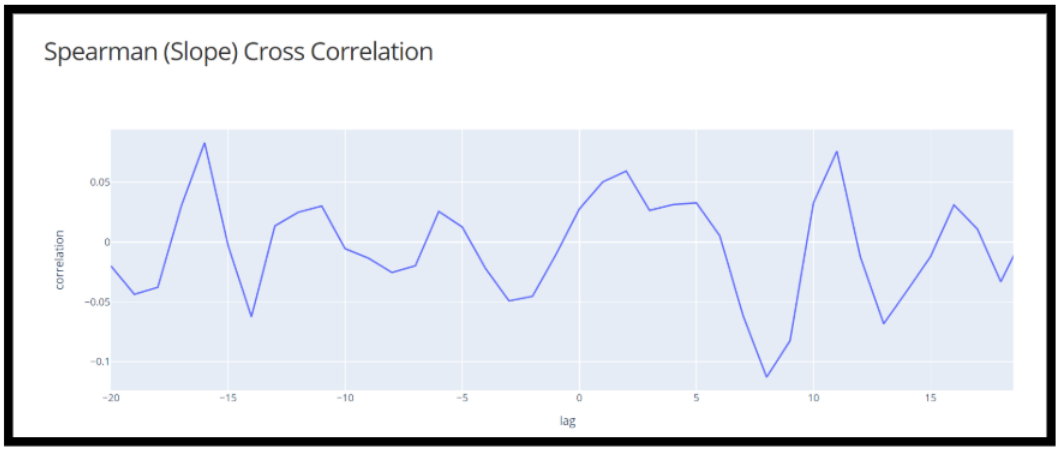

Figure 9: Cross correlation using Spearman correlation method

5.4.2 Pearson Correlation

Pearson is one of the most widely used correlation approaches. It is for variables with a linear relationship and data being normally distributed. Linearity and homoscedasticity are also assumed in this statistical correlation method. The relationship of the sentiment score is found to be positive using Pearson statistical method. The negative correlation is observed to be -0.04 at lag 10 and highest correlation is observed as 0.07 at lag -18.

Figure 10 shows the cross correlation between the derived sentiment score from the tweets and the bitcoin price using Pearson correlation method.

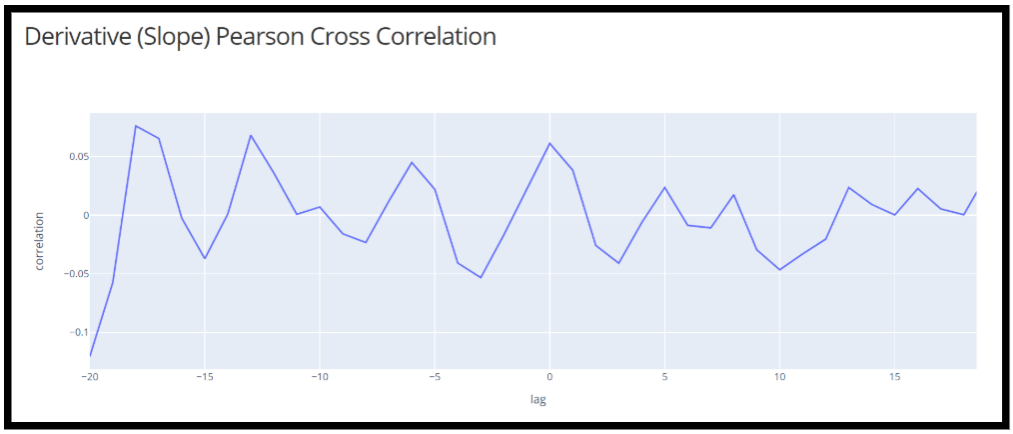

Figure 10: Cross correlation using Pearson correlation method

5.4.3 Kendall Correlation

Kendall Correlation is a non-parametric statistical technique that measures the strength of dependency between two or more variables, similar to Spearman Rank Correlation. The maximum correlation is observed to be 0.049 at lag 11 and the negative correlation is observed to be -0.07 at lag 8.

Figure 11 shows the cross correlation between the derived sentiment score from the tweets and the bitcoin price using Kendall correlation method.

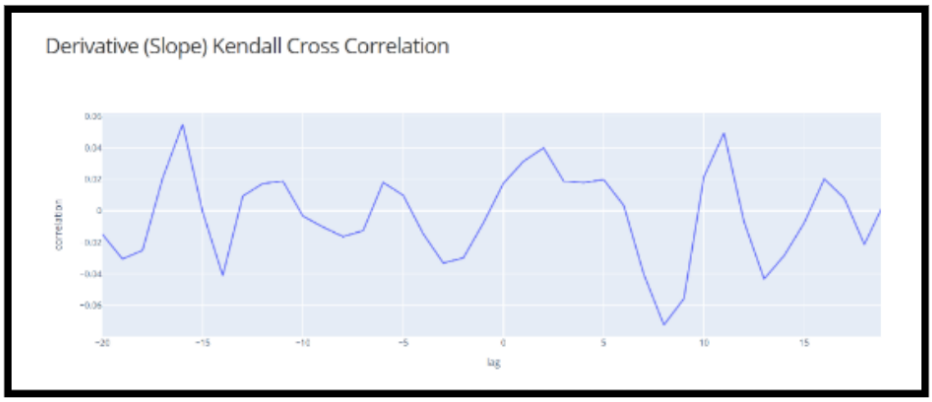

Figure 11: Cross correlation using Kendall correlation method

5.5 Prediction Model

Long Short-Term Memory (LSTM) is similar to recurrent neural network that was created to address long-term reliance issues. Unlike RNN, LSTM features structured gates that control the cell state. The sigmoid layer aids in the control of this gate, which in turn controls the cell state. This allows LSTM to account for both short and long-term information passing through these gates. The model is built using the python libraries sklearn and keras. The data is first normalized and then split into train and test data. 0.9 of the data is split as train and 0.1 of the data is split as test data. The training model was then reshaped and executed with 500 epochs. Two optimizers - RmsProp and Adam are used in the analysis.The predicted values using each optimizer is plotted in a graph. There performance is compared and discussed in the next section.

Figure 12 explains the LSTM network.

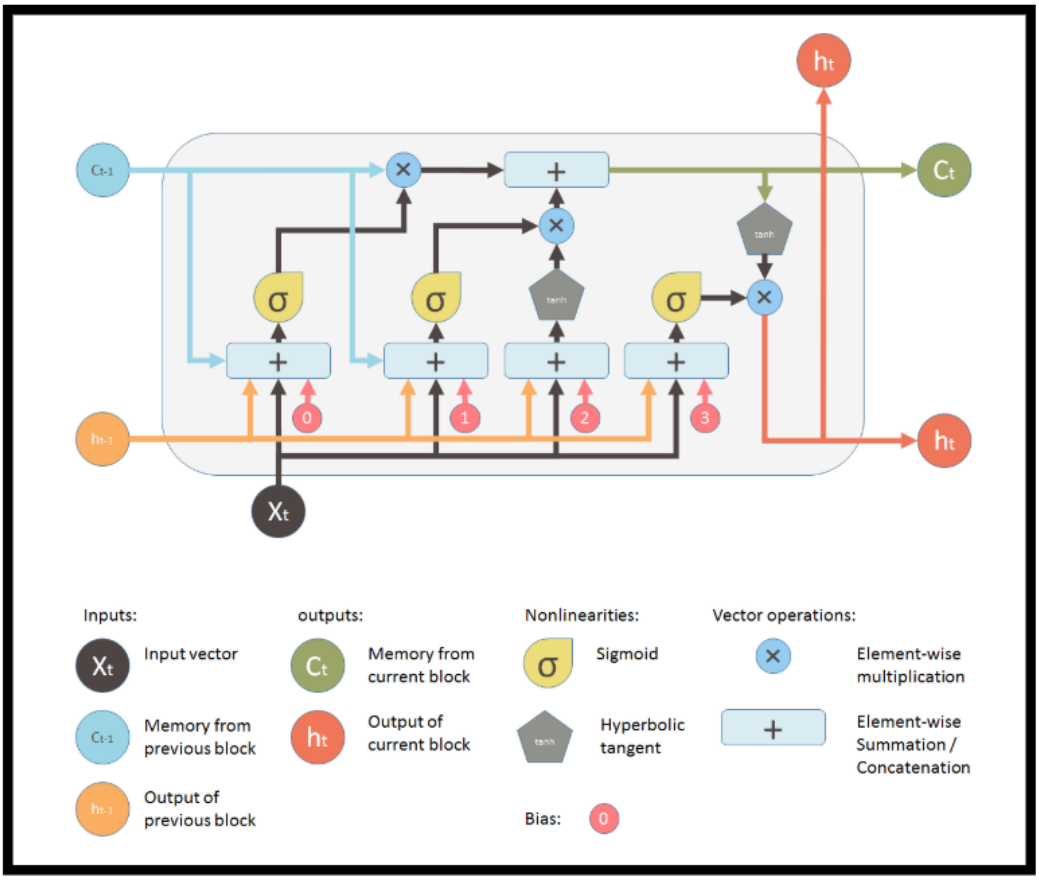

Figure 12: LSTM Network

6 Evaluation, Results and Comparison of Developed Prediction Models

The bitcoin price is predicted based on the derived score using the LSTM model and the results for both the optimizers against the actual values are presented in Figure 13. The direction and the closing price of the bitcoin are the main objectives of the project. Due

to the high volatility of the bitcoin which is caused by many other factors and considering public opinion to be one of the impacting factors, the exact price of the bitcoin was not resulted considering the sentiment score as an only factor. However, the direction of the bitcoin price seems to match quite well with the actual bitcoin price. It is observed that Adam optimizer had performed better in terms of the price prediction in comparison to Rmsprop. While Rmsprop performed well in the change in direction of the bitcoin price.

Figure 13 shows the output of LSTM model with predicted bitcoin price.

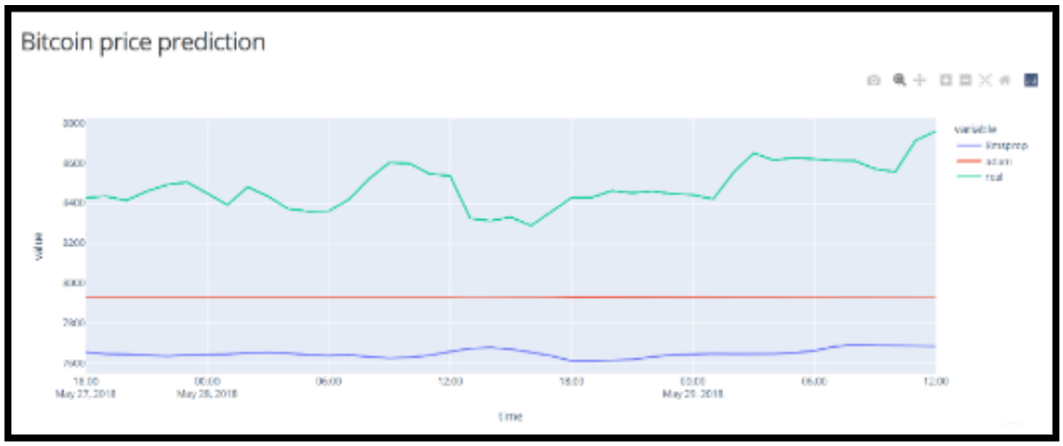

Figure 13: Output of LSTM model with predicted bitcoin price

To evaluate this predicted price of the bitcoin using the LSTM Model metrics like Root Mean Squared Error, Mean Squared Error, Mean Absolute Percentage Error, Mean Tweedie Deviance, Mean Gamma Deviance and Mean Poisson Deviance are used. Each metric is explained in the next sections.

6.1 Mean Squared Error

Mean Squared Error(MSE) is the average of the square of the difference between the actual value and predicted value. It is used as a measure to evaluate models that deal with regression problems. For the LSTM model with Rmsprop as an optimizer, The Mean Squared Value was observed as 710265. While it is 318483 using Adam as an optimizer. Low value of MSE indicates a better model. The LSTM model Adam was an optimizer outperformed LSTM model with Rmsprop as an optimizer considering the Mean Squared Error metric.

6.2 Root Mean Squared Error

Root Mean Squared Error is one of the most used evaluation metric for getting the error rate of the prediction models. It conveys the distance of the residuals from regression data line.This validation metric squares residuals after describing their dispersion around the line of best fit. It is important for analyzing and minimizing regression model errors. The RMSE value for the LSTM model to predict the bitcoin price using Rmsprop is 842. While for the LSTM Model built using Adam as an optimizer, the RMSE is 564. Lower value of RMSE indicates a better model. Considering the RMSE metric, the LSTM model

designed using the Adam optimizer outperformed the LSTM model with Rmsprop as an optimizer.

6.3 Mean Absolute percentage Error

MAPE, which stands for mean absolute percentage error, is one of the most prominent metrics used to evaluate a model's forecasting accuracy. It is widely used as it is easy to explain and easy to interpret. The lower MAPE value indicates a better predictive model. For the LSTM model designed using Rmsprop as an optimizer, the MAPE value is 9.8 percentage. While the LSTM model designed using Adam as an optimizer attained a MAPE value is 6.5 percentage. This indicates that the Adam optimizer had predicted price more accurately than Rmsprop.

Figure 17 shows the formula of MAPE calculation.

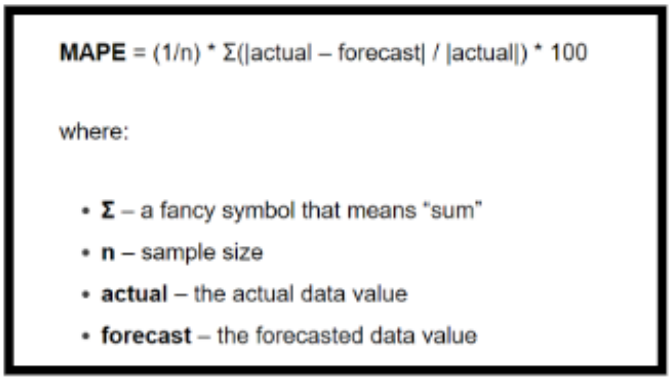

Figure 14: MAPE calculation formula

6.4 Mean Tweedie Deviance

With a power argument, the mean tweedie deviance function calculates the Tweedie deviance error (). This is a metric that obtains the regression targets' projected expectation values.

$$\mathrm{D}(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \begin{cases} (y_i-\hat{y}_i)^2, & \text{for } p=0 \text{ (Normal)} \\ 2(y_i \log(y/\hat{y}_i)+\hat{y}_i-y_i), & \text{for} p=1 \text{ (Poisson)} \\ 2(\log(\hat{y}_i/y_i)+y_i/\hat{y}_i-1), & \text{for} p=2 \text{ (Gamma)} \\ 2\left(\frac{\max(y_i,0)^{2-p}}{(1-p)(2-p)}-\frac{y\,\hat{y}_i^{1-p}}{1-p}+\frac{\hat{y}_i^{2-p}}{2-p}\right), & \text{otherwise} \end{cases}$$

Figure 15: Tweedie calculation formula

When p=0 in the above formula, the first degree difference between the actual value and the predictive value is quantified which is similar to that of MSE. As discussed above

considering MSE, LTSM model built with Adam as an optimizer outperformed the LSTM model with Rmsprop as an optimizer.

6.5 Mean Poisson Deviance

When p value in the formula shown in Figure 17 is 1, then the metric is called Mean Poisson Deviance. The deviation scales linearly for Poisson distributions with power=1, and quadratically for Normal distributions with power=0. Extreme discrepancies between true and anticipated targets are given less weight as power increases.The bitcoin prediction LSTM model with Rmsprop as an optimizer has a mean Poisson Deviation of 89.5, while the model with Adam as an optimizer has a mean Poisson Deviation of 39.1.

6.6 Mean Gamma Deviance

When p value in the formula shown in Figure 17 is 2, then the metric is called Mean Gamma Deviance. It means that the scaling of ytrue and ypred has no effect on the deviance. The Mean Gamma Deviation for the bitcoin prediction LSTM model with Rmsprop as an optimizer is 0.01 and for the model with Adam as an optimizer, it is 0.004.

6.7 Discussion

Evaluation metrics help in assessing the correctness and efficiency of the predictive model. Mean Squared Error, Root Mean Squared error, Mean Absolute percentage Error, Mean Tweedie Deviance, Mean Poisson Deviance and Mean Gamma Deviance are used to evaluate the LSTM model built to predict the bitcoin price based on the sentiment score derived from the tweets.

Figure 16 shows the output of LSTM model using Rmsprop optimizer.

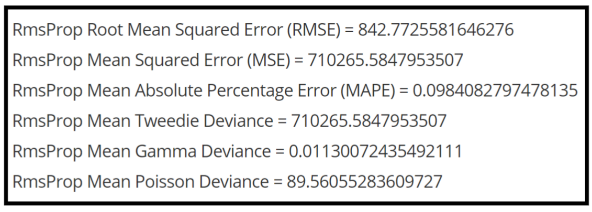

Figure 16: Output of LSTM model using Rmsprop optimizer

Figure 17 shows the output of LSTM model using Adam optimizer.

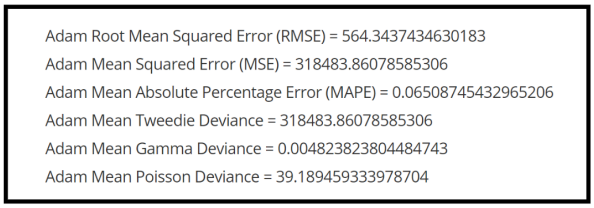

Figure 17: Output of LSTM model using Adam optimizer

The LSTM model with Adam as an optimizer decisively outperformed the LSTM model with Rmsprop as an optimizer on every evaluation criteria.

7 Conclusion and Future Work

The project's objectives have been achieved, and the research question posed have been answered. The correlation coefficients determined using several statistical techniques such as Spearman, Pearson, and Kendall correlation coefficients were used to accomplish this. The favorable association and strength of the relationship were discovered by this experiment. The implementation of Bitcoin prediction based on the sentiment score derived from the tweets is achieved using the LSTM model. Two price prediction is done based on two optimizers - Rmsprop and Adam. Where the model built using Adam as an optimizer outperformed Rmsprop.

The core objective of the project was to determined the relation between the change in opinion of the people with the change in the bitcoin price. The tweets were extracted with the tweeter API and the sentiment score was calculated using VADER algorithm which was used as main factor to predict the bitcoin price for next hour. Also, the code built is flexible to predict live bitcoin price based on the last week's data. The relationship between the sentiments determined from twitter and bitcoin price was observed to be strong and positive.

To construct stronger models, further study on this paper may consider incorporating additional variable values and expanding the historical span of time. Also, bitcoin entails in large investments, accurate prediction of the bitcoin price can be done considering other impacting factors.

8 Acknowledgement

I would like to express my gratitude to Dr. Catherine Mulwa, my supervisor, for her mentorship and guidance in completing my research project on time. She has always offered me advice that have greatly aided me in overcoming the obstacles I encountered during my dissertation.

References

Ahn, Y. and Kim, D. (2020). Emotional trading in the cryptocurrency market, *Finance Research Letters* p. 101912.
URL: *https://www.sciencedirect.com/science/article/pii/S1544612320317268*

Al-Ghuribi, S. M., Mohd Noah, S. A. and Tiun, S. (2020). Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews, *IEEE Access* **8**: 218592–218613.

Balfagih, A. M. and Keselj, V. (2019). Evaluating Sentiment C1assifiers for Bitcoin Tweets in Price Prediction Task, *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5499–5506.

Bibi, S., Hussain, S. and Faisal, M. I. (2019). Public Perception Based Recommendation System for Cryptocurrency, *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 661–665. ISSN: 2151-1411.

Gualberto, E. S., Sousa, R. T. D., Vieira, T. P. D. B., Costa, J. P. C. L. D. and Duque, C. G. (2020). From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection, *IEEE Access* **8**: 76368–76385. Conference Name: IEEE Access.

Han, W., Tian, Z., Zhu, C., Huang, Z., Jia, Y. and Guizani, M. (2021). A Topic Representation Model for Online Social Networks Based on Hybrid Human–Artificial Intelligence, *IEEE Transactions on Computational Social Systems* **8**(1): 191–200. Conference Name: IEEE Transactions on Computational Social Systems.

Martin, K., Alsmadi, I., Rahouti, M. and Ayyash, M. (2020). Combining Blockchain and Machine Learning to Forecast Cryptocurrency Prices, *2020 Second International Conference on Blockchain Computing and Applications (BCCA)*, pp. 52–58.

Narman, H. S. and Uulu, A. D. (2020). Impacts of Positive and Negative Comments of Social Media Users to Cryptocurrency, *2020 International Conference on Computing, Networking and Communications (ICNC)*, pp. 187–192. ISSN: 2325-2626.

Saadah, S. and Whafa, A. A. A. (2020). Monitoring Financial Stability Based on Prediction of Cryptocurrencies Price Using Intelligent Algorithm, *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 1–10.

Sun, J., Zhou, Y. and Lin, J. (2019). Using machine learning for cryptocurrency trading, *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, pp. 647–652.

Yao, W., Xu, K. and Li, Q. (2019). Exploring the Influence of News Articles on Bitcoin Price with Machine Learning, *2019 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6. ISSN: 2642-7389.

Yiying, W. and Yeze, Z. (2019). Cryptocurrency Price Analysis with Artificial Intelligence, *2019 5th International Conference on Information Management (ICIM)*, pp. 97–101.

Zhang, W., Yoshida, T. and Tang, X. (2008). TFIDF, LSI and multi-word in information retrieval and text categorization, *2008 IEEE International Conference on Systems, Man and Cybernetics*, pp. 108–113. ISSN: 1062-922X.