MCAST

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

Social Media Sentiment Analysis for Cryptocurrency Price Prediction

Mr. Ryan Pirotta

Mentored by Mr. Marco Farrugia

June 2022

A dissertation submitted to the Malta College Arts Science and Technology (MCAST) in partial fulfilment of the requirements for the degree of B.Sc. (Hons.) in Software Development.

Authorship Statement

This dissertation is based on the results of research carried out by myself, is my own composition, and has not been previously presented for any other certified or uncertified qualification.

The research was carried out under the supervision of Mr. Marco Farrugia.

Signed: Ryan Pirotta *Date: 06/06/2022*

Copyright Statement

In submitting this dissertation to the MCAST Institute of Information and Communication Technology. I understand that I am giving permission for it to be made available for use in accordance with the regulations of MCAST and the Library and Learning Resource Centre. I accept that my dissertation may be made publicly available at MCAST's discretion.

Signed: Ryan Pirotta *Date: 06/06/2022*

Abstract

Over the last decade, institutions have demonstrated interest in the fundamentals of Bitcoin. Due to its fixed supply, this cryptocurrency becomes a store of value – with some arguing that it is a form of digital gold. An interesting aspect of the cryptocurrency market is price volatility, mainly since the price of an asset relies on people's perceptions and opinions.

This study aimed to predict the next day's price direction of Bitcoin using social media sentiment analysis. The price direction was measured by taking the current closing price with the previous day, where a value of 0 and 1 was given whether it was down or up respectively. Therefore, this research was classified as a binary classification problem.

As social media platform is widely used for sharing opinions about any topic, Twitter was used as a measure to predict the price of Bitcoin in this research through Natural Language Processing (NLP). The tweets undertook several cleaning processes as social media content contains an abundance of noise. Two different sentiment analysers, VADER and TextBlob are used to extract the polarity from the Twitter posts. An experiment was conducted on an annotated dataset provided by NLTK Library to analyse which cleaning process is more suitable for these particular sentiment analysers on social media content. Multivariate forecasting was introduced using different lagged observations to enhance the model results. Different lagged observations were introduced such as 1, 3 and 7-day lag. To predict the next day's price movement of Bitcoin, two different neural network models, Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) were used in order to predict the Bitcoin price movement. An ensemble learning method was taken into consideration using Random Forest (RF) classifier. Given certain correlation results, the dataset was further trimmed, and models were executed for a final time to ensure the optimal result.

Results have demonstrated that the LSTM using TextBlob features achieved the highest maximum accuracy of 64.58% and an overall accuracy of 63.21% with a mean F1-Score of 63.06%. This research proved that sentiment analysis combined with deep learning models or ensemble learning methods can predict the price direction to a certain extent. Moreover, sentiment tended to be more positive regardless of the price falling. This concluded that people tend to have a conflict of interest towards the asset.

Acknowledgements

I would like to express my sincere gratitude to the kind and helpful people who have offered me their assistance at every stage of the research project.

Primarily, I would like to thank my supervisor, Mr. Marco Farrugia, for his constant support and mentorship. Secondly, I would like to thank my family, without whom this journey would have not been possible.

Contents

List of Abbreviations

ANEW – Affective Norms of English Words - 8

BERT – Bidirectional Encoder Representation from Transformer - 10

BiLSTM – Bidirectional Long Short-Term Memory - 2, 5, 10, 15, 16, 28, 29, 37, 38, 41, 42, 44, 45, 46, 54

CNN – Convolutional Neural Network - 15

CSV – Comma Separated Values – 54

GI – General inquirer - 8

GLUE – General Language Understanding Evaluation - 10

LIWC – Linguistic Inquiry and Word Count - 8

LSTM – Long Short-Term Memory - 2, 5, 9, 10, 14, 15, 16, 28, 29, 37, 38, 39, 41, 42, 44, 45, 46, 47, 54

MLP – Multi-layer Perceptron – 15, 16, 17

NLP – Natural Language Processing - 1, 7, 9, 10, 21, 46, 54

NLTK – Natural Language Tool Kit - 2, 9, 21, 22, 23, 24, 30, 32, 33, 54

NSP – Next Sentence Prediction - 10

RBFNN – Radial Basis Function Neural Network - 14

RF – Random Forest - 2, 5, 8, 9, 12, 15, 27, 28, 30, 36, 37, 38, 40, 41

RNN – Recurrent Neural Network - 9, 10, 16

SQuAD – Stanford Question Answering Dataset - 10

SWAG - Situations with Adversarial Generations - 10

VADER - Valence Aware Dictionary and Sentiment Reasoner - 2, 3, 8, 9, 13, 14, 21, 24, 25, 26, 32, 33, 34, 35, 36, 37, 39, 40, 41, 43, 45, 46, 54

List of Figures

List of Tables

Chapter 1. Introduction

The title “Social Media Sentiment Analysis for Cryptocurrency Price Prediction” defines the intention of the present research. Sentiment Analysis is a subfield of Natural Language Processing (NLP) which analyses textual content and classifies whether it is positive, neutral, or negative. This can be used as guidance to know the user sentiment on the investigated scenario, in this case, to indicate future prices. On the other hand, cryptocurrencies are a type of medium of exchange to buy goods, like any other traditional currency. Albeit cryptocurrency tend to be highly volatile in nature, this can be a good opportunity for traders or investors to make a profit.

The largest cryptocurrency in terms of market capitalization, Bitcoin was founded in November 2008 by an anonymous person or group of people under the pseudonym Satoshi Nakomoto. On the 3rd of January 2009, Bitcoin was launched publicly. The underlying technology is considered a decentralized peer-to-peer payment mechanism, making it a medium of exchange without needing any intermediary third-party to perform a transaction since it is based on a peer-to-peer network. In 2017, the price of Bitcoin increased remarkably by 2000%, which attracted a global interest into cryptocurrencies. Due to its distinctive characteristic of decentralization when compared to traditional Fiat currencies, Bitcoin and cryptocurrency in general, are one of the most intriguing forms of digital transactions worldwide. The price fluctuations rely on the people’s changing perceptions and opinions rather than following institutional regulations. This also means that the digital currency values are highly volatile, which makes it difficult to be used as a global currency (Wołk, 2020). Moreover, Bitcoin is characterized as a store of value due to its fixed max supply and deflationary design (Baur and Dimpfl, 2021).

In this day and age, there are multiple platforms available for people to voice their opinions. Twitter is one of the most widely used social media platforms, which collects multidimensional perspectives and people's perceptions worldwide (Wołk, 2020). This study utilised Twitter as the preferred marketing tool for cryptocurrency topics to use as a means to foresee their price movement. According to Karalevicius et al. (2018), sentiment analysis technique can be efficient in predicting the price direction of cryptocurrencies. Thus, this study investigated the correlation of Twitter sentiment towards the price direction of Bitcoin.

The main research questions investigated in this study are:

RQ1: Can the next day's cryptocurrency price direction be predicted by utilising social media sentiment analysis?

To tackle the first research question, the extracted sentiment from Twitter and the Tweet volume, as suggested by Abraham et al. (2018), was used as an indicator. Moreover, Random Forest (RF) classifier and Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) networks were utilised as deep learning models to predict the price direction of Bitcoin based on Twitter sentiment.

RQ2: Does VADER outperform TextBlob sentiment analyser when extracting the polarity from social media posts?

Pre-trained sentiment analysers VADER and TextBlob are both attuned to social media content. These two sentiment analysers were tested out on a dataset provided by Natural Language ToolKit (NLTK) library. Different combinations of cleaning processes were performed to understand better the sentiment analyser performance on social media content and the confirmation of which achieved the best overall accuracy for extracting polarity.

RQ3: Does social media Sentiment Analysis provide an efficient indicator for cryptocurrency price prediction?

Previous studies in this field had excluded sentiments as an indicator for price prediction as these were said to be more subjective due to conflict of interest (Abraham et al., 2018; Wołk, 2020). Thus, this study will attempt to verify whether sentiment is an efficient indicator when utilising TextBlob and VADER sentiment analysers.

RQ4: Which model from Long Short-Term Memory, Bidirectional Long Short-Term Memory and Random Forest Classifier provides the best performance in predicting the next day's cryptocurrency direction?

After thorough research related to cryptocurrency prediction, it was found that, although related studies achieved optimal results, different data represented different anomalies for models predicting prices using sentiment analysis on Twitter data. Some of these anomalies were due to the evaluation of small historical datasets and thus lacked detailed analysis of the quality of the content. Challenges arose from certain issues mentioned as predicting price direction is not straightforward to solve.

A Twitter post is a sentiment of a user which might influence other users. Twitter bots are used for marketing purposes and certain tweets might be duplicated (Valencia et al., 2019). Sentiment analysis requires one to consider that social media platforms including Twitter contain features such as hashtags, profile mentions and hyperlinks which may result in noise. In addition, certain Tweets tend to be highly subjective since users are inclined to favour a profitable outcome for themselves resulting in conflicts of interest.

In the next chapter, background information about the terminology, concepts found in this research and preceding related work are discussed. The following chapter is the research methodology which is followed by the discussion of the results chapter which will present the findings through certain methodologies that this study undertook. The final chapter gives a conclusion with a short recap and discussion of the outcomes, limitations and possible enhancements for future work, including a summary of this research project.

Chapter 2. Background and Literature Review

This chapter highlights all relevant background information and other studies related to the prediction of cryptocurrency price direction using Sentiment Analysis and machine learning algorithms. Some features covered include Cryptocurrency, Twitter, Sentiment Analysis, Polarity Classification, Multivariate Forecasting, RF Classifier, and different neural networks such as LSTM and BiLSTM.

2.1 Background Information

The performance and direction of cryptocurrencies performance are influenced by several factors. These include the utility of the currency itself and the influence on social media platforms such as Twitter or YouTube. The influence of the latter can be analysed by employing sentiment analysis techniques together with artificial intelligence.

2.1.1 Cryptocurrency

A cryptocurrency is a form of a digital asset based on a network that is distributed across a large number of computers. The word ‘cryptocurrency’ consists of two parts: ‘crypto’ and ‘currency’: ‘Crypto’ reflects the security of the asset since it is based on cryptography, which makes it difficult for it to be altered or hacked; ‘Currency’ is because this asset is also a medium of exchange (Abd Aziz et al., 2022). Introduced in the market in 2009, Bitcoin is the first cryptocurrency invented by an anonymous individual or a group of individuals under the pseudonymous Satoshi Nakamoto (sibel Kervanci and Fatih, 2020).

The fundamentals of Bitcoin are defined in the white paper of Bitcoin where Nakamoto (2008) highlights that Bitcoin is based on a peer-to-peer network, thus making it a decentralized digital currency where there is no need for any intermediary third-party to perform a transaction. To make this possible, cryptocurrencies are based on an underlying technology known as ‘Blockchain’, which utilises a tamperproof and transparent ledger, whose transactions are immutable (Poongodi et al., 2020). Blockchain is the system that brought cryptocurrency into existence due to its decentralization and security. Bitcoin is considered a store of value since it has a fixed supply and a deflationary design (Baur and Dimpfl, 2021).

Blockchain technology is decentralised, so it caters for issues of security and trust found in traditional currency in several ways. The new blocks that are appended to the blockchain are stored linearly and chronologically, where every block is appended at the end of the chain. Every node in the peer-to-peer network has a copy of the ledger consisting of blocks with chronologically created transactions, each of which created by using cryptography to link together newer blocks with the previous block. A modification in an older block requires updating the subsequent blocks making it difficult to hack.

Bitcoin is the largest cryptocurrency in terms of market capitalisation, having a current market cap of around 1 trillion dollars (CoinMarketCap, 2021). Bitcoin and cryptocurrency, in general, are one of the most intriguing forms of digital transactions around the globe (Nakamoto, 2008). The number of cryptocurrencies is increasing relentlessly since the invention of Bitcoin, over 12,000 cryptocurrencies have been created (CoinMarketCap, 2021). Moreover, traders can make huge profits from investing into cryptocurrencies due to their high volatility.

2.1.2 Twitter

In the online era, various channels are widely available for people to voice their opinions. Twitter is one of the most widely known social media platforms which gathers multidimensional perspectives and people’s perceptions worldwide (Wołk, 2020). According to Alothali et al. (2018) in the third quarter of 2007, Twitter had 330 million active users. By 2015, Twitter's active users had grown to an estimated 1.3 billion.

Twitter’s existence dates back to July 2006 as an application in both the social media space and microblogging. Microblogging is a type of medium, similar to a blog but allows for more frequent and smaller posts (Abraham et al., 2018). These posts are called ‘tweets’ and they can be sent and made public online. Originally, these posts were 140 characters long, but since November 2017 the length has doubled down to 280 characters. Furthermore, users on Twitter can reply to a tweet and even reshare a tweet posted by another user. Moreover, since the launch of Twitter certain features such as hashtags and profile mentions were introduced. Hashtags ‘#’ allow categorization of tweets while profile mentions ‘@’ links a user to a Twitter post. Twitter gained its popularity for the rich source of data on how people feel about nearly any given topic, including within the various crypto communities. It should be noted that according to Alothali et al. (2018), between 9% and 15% of users are bots which accumulate to an equivalent of 41 million users. These bots and their activities may hinder the performance of sentiment analysis.

2.1.3 Sentiment Analysis

Sentiment analysis is a subdivision of NLP and it can be used in various areas such as Information Extraction, Email Spam Detection, Machine Translation, Summarization and so on (Khurana et al., 2017). The objective of sentiment analysis is to extract subjective information from textual documents whether it be in the form of reviews, emails, news articles, tweets, and so on, to get a better understanding of the human language (Abraham et al., 2018). The combination of machine learning and digital datasets enhances recognition in this area of study. Abraham et al. (2018) highlighted that Twitter’s data is widely used to extract and analyse the sentiment of the tweets. A typical task of sentiment analysis is polarity classification where polarity is extracted from a textual context (Yue et al., 2019).

2.1.4 Polarity Classification

Polarity classification is the process of annotating a textual document such as reviews and news articles, and classifying them into categories such as positive, negative, or neutral. Polarity lies in the range of (-1 to 1) where in a 5-class tier -1 stands for strongly negative, -0.5 for negative, 0 for neutral, 0.5 for positive and 1 stand for strongly positive. Whilst 3-class tier -1 is negative, 0 is neutral and 1 is positive (Yue et al., 2019). In a 2-class tier, 1 stand for positive and -1 for negative excluding the neutral category.

Prakash and Aloysius (2019) mentioned that there are two techniques to extract the polarity from a phrase, which are the Machine Learning or the Lexicon-based approach. The Machine Learning approach includes supervised and unsupervised algorithms where the input needs to be fed to these types of algorithms, which requires training and testing of the dataset. The training part is where the document is learnt by the system, and the testing part is the validation performance (Prakash and Aloysius, 2019). Moreover, the latter study mentioned that the lexicon-based approach has two types of classification which are Dictionary-based and Corpus-based. The Dictionary-based approach consists of collecting data manually and its information is used to search for synonyms and antonyms from WordNet and sentWordNet dictionaries. The Corpus-based approach focuses on dictionaries related to a specific domain where words are related to semantic and statistical methods such as LSA (Prakash and Aloysius, 2019). VADER is a sentiment analyser that follows a Lexicon and Rule-based approach that is particularly attuned to sentiments in social media. VADER is an open-source application, built on three validated lexicon features such as LIWC, ANEW and GI (Bonta et al., 2019). The benefits of VADER are not only classifying text as positive, neutral, or negative but also measuring the intensity of the words utilised (Abraham et al., 2018).

Another popular sentiment analyser is TextBlob. It is a Python library for processing textual data, and it is built upon the NLTK library which provides a consistent API for common NLP tasks. This model is a Lexicon and Rule-based approach similar to the VADER analyser. It has multiple functionalities such as Noun phrase extraction, part-of-speech tagging, Sentiment Analysis, Language Translation and detection, n-grams and spelling correction. The sentiment analysis operation provides the polarity and subjectivity of the textual content (Bonta et al., 2019).

2.1.5 Random Forest Classifier

RF classifier is an ensemble learning method for classification that contains several decision trees, which are the building blocks of the RF model (Akyildirim et al., 2021). The RF classifiers fall under many subtopics or categories of ensemble-based learning models. Consequently, the RF resides in a large number of trees that operate as an ensemble. It is efficient in operation, easy to implement and also proved to be efficient in particular domains. Furthermore, RF functions in a manner that creates a set of decision trees from the subset of the training set and each tree makes a prediction.

2.1.6 Long Short-Term Memory Networks

Recurrent Neural Network (RNN) is a class of artificial neural networks, where the connection between nodes transforms an undirected or directed graph with a sequential time series. In traditional RNNs, where the network is trained through backpropagation through time, a common problem known as exploding and vanishing gradients occurs (Pienaar and Malekian, 2019) when a large error gradient starts accumulating.

Long Short-Term Memory (LSTM) network was invented with the intention to overcome these issues that occur in RNN architectures (Pienaar and Malekian, 2019). LSTMs are a type of artificial RNN architecture that is utilised in deep learning. LSTM can process both single data points, such as images and sequences of data, such as speech, video, human activity and so on. Each LSTM unit consists of a cell, and three gates which are called input, output, and a forget gate. Each cell remembers values of arbitrary time intervals, while the gates regulate the information flow both from and to the cells (Siami-Namini et al., 2019). LSTM is ideal for classification, processing and predicting time-series based data and it is also proven that it excels in processing, learning, and classifying such types of data. RNNs and LSTMs are very similar, with the only difference that LSTMs have a hidden third layer that contains memory blocks with cells that can store information over a long period, while traditional RNNs cells contain a single internal layer that acts on the current state and input (Ferdiansyah et al., 2019).

2.1.7 Bidirectional Long Short-Term Memory Networks

Bidirectional Long Short-Term Memory (BiLSTM) is an extension of the LSTM in which two LSTM models are applied. The first model (e.g., forward layer) memorises the sequence of the inputted data provided, and the second model (e.g., backward layer) memorises the sequence in backwards direction. The BiLSTM results in an improvement in long-term dependencies since the model is applied twice, consequently obtaining a model with better accuracy (Siami-Namini et al., 2019). The Bidirectional Encoder Representations for Transformers (BERT) language model was developed by Google, and it is based on bidirectionality and pre-training. BERT is a transformer-based machine learning technique for NLP which results to be useful since it can extract details about a word or phrase sequence from both directions. In the NLP field, BERT is known for the state-of-the-art performance of certain tasks such as GLUE, SWAG, SQuAD, Next Sentence Prediction (NSP), etc… (Devlin et al., 2018).

2.2 Related Work

This chapter gives an overview of other people's work that is related to cryptocurrency predictions and their approaches for prediction including data collection, sentiment analysis and different neural networks.

2.2.1 Data Collection

One of the main pillars for predictions revolves around data. Nowadays, collecting data is done with such ease and efficiency through the assistance of the Internet. Balfagih and Keselj (2019) and Iqbal et al. (2021) used Kaggle for data gathering related to the cryptocurrency prices and collected individual posts about opinions about that particular cryptocurrency in the same period. Furthermore, Kraaijeveld and De Smedt (2020), mentioned that some researchers in other studies utilised Reddit and Bitcointalk as social media platforms to collect users' sentiments. A study by Abraham et al. (2018) utilised Google Trends to obtain a better idea of the users' trends over a particular period. Google Trends provides the number of searches performed on Google and that is linked to a particular topic, which returns a Search Volume Index.

Despite this, most of the studies related to the use of sentiment analysis to predict the price of cryptocurrency use Twitter. Twitter data can be collected by using crawlers based on the API or the TwitterAPI itself (Abraham et al., 2018; Balfagih and Keselj, 2019; Jain et al., 2018; Kilimci, 2020; Mohapatra et al., 2019; Pagolu et al., 2016; Rathan et al., 2019). Moreover, the data needed relating to financial information can be collected using several sources that provide an API, such as coinmarketcap.com, coindesk.com, cryptocompare.com, quandl.com and finance.yahoo.com (Jain et al., 2018; Mohapatra et al., 2019; Pagolu et al., 2016; Rathan et al., 2019; Stenqvist and Lönnö, 2017). The most common fields collected from Twitter consist of the actual text of the post, the username, and the number of likes and retweets while the common fields of financial datasets are the opening, closing, lowest, and highest prices, trading volume and transactions in a particular timeframe.

2.2.2 Data Pre-processing and Cleaning

This section highlights the importance of pre-processing and data cleaning. The data gathered from social media platforms contain an abundance of noise which may influence the performance of the sentiment analysis. Hence, in order to prepare the data to perform sentiment analysis, it is necessary to undertake a cleaning process. According to the aforementioned studies this process includes some common steps such as the removal of duplicate, irrelevant and non-English tweets (Kraaijeveld and De Smedt, 2020; Stenqvist and Lönnö, 2017). Moreover, a popular phase is the Tokenisation where phrases are divided into arrays to analyse words at ease. Once words are tokenised, the next phase is Stop Words Removal, where a set of commonly used words are eliminated since they have minor useful information, including common words like ‘a’, ‘is’, ‘want’ and so on (Kraaijeveld and De Smedt, 2020).

The tokenisation process makes it easier to eliminate capitalisation, hyperlinks, and emoticons (Balfagih and Keselj, 2019). Symbols such as hashtags (#), question marks, quotes, mentions (@) and others are also removed as it may influence the analyser’s performance (Kraaijeveld and De Smedt, 2020; Stenqvist and Lönnö, 2017). The latter paper mentions the importance of removing extra white spaces, numerical characters, and phrases with less than 4 words since they are not suitable for sentiment analysis (Kraaijeveld and De Smedt, 2020). Also, Twitter contains an abundance of tweets that are generated by bots which may lead the analysis off track.

2.2.3 Sentiment Analysis

Sentiment analysis is used to extract the sentiment polarity from the data, which must be performed on a cleaned input, using the previous steps discussed to clean the data. Various sentiment analysis techniques exist to extract the polarity of an opinion. This can be done by utilising software packages that generate the polarity classification and the subjectivity score.

Word Embedding models such as the Bag-of-Words is capable of representing text into numbers or vectors since the machine can only understand 1's and 0's. In a study by (Pant et al., 2018), two different techniques were used for features extraction, namely the Bag-of-Words and Word2Vec. In addition, to obtain the utmost performance out of the features extracted, a voting classifier was created with the purpose to choose the highest sentiment score among five different algorithms such as Naïve Bayes, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Linear Support Vector Classifier and RF. In the latter paper, the Bag-of-Words performed better than the Word2Vec model.

Research by Pano and Kashef (2020), Rathan et al. (2019) and Valencia et al. (2019) extracted the polarity using the VADER sentiment analyser and highlighted its benefits, mainly due to it being human validated, open-source and it was developed with the purpose of being used on social media content such as Twitter. Apart from these advantages mentioned, VADER label's each input or opinion as positive, neutral, and negative in addition to an overall score, better known as a compound score. The compound score entitles the summation of the input scores.

A study by Jain et al. (2018) and Gurrib and Kamalov (2021), used another popular sentiment analyser named TextBlob to extract the polarity of the tweets. TextBlob provides the polarity of a textual content and the subjectivity of it as well. Moreover, Linardatos and Kotsiantis (2020) used both VADER and TextBlob as their sentiment analysers to predict Bitcoin price movement.

A research conducted by Wooley et al. (2019) highlighted that in the world of cryptocurrency, certain individuals with high profiles have more influence on the market so some additional weight needs to be given to them. This can be done by utilising the PageRank algorithm or a personalised one (Wooley et al., 2019). Thus, users or classified tweets are subjected to the PageRank algorithm to make the sentiment more effective. A survey by Riquelme and González-Cantergiani (2016) mentioned multiple approaches including topics sensitive to rank users' accounts using different features such as the follow-up relationship, retweets, likes, replies, timeline and content analysis.

2.2.4 Lagged Features

Certain sentiments might take effect on the market with a delay. A study by Hotz-Behofsits et al. (2018) emphasize the importance of introducing a wide range of univariate and multivariate models for time series data to predict the next day's price of Bitcoin. Another study by Jana et al. (2021) utilized a multivariate forecasting, introducing from 1 to 5-lag days to predict the next day's price. In a study by Mudassir et al. (2020) focused on the end-of-day closing price prediction and the direction in price prediction. These predictions were made for short-term calls such as daily closing forecasting and for the mid-term which ranges from 7 days to 90 days. Moreover, the latter study achieved an accuracy of the best model of 64% for the 90th day using LSTM neural network.

2.2.5 Price Predictions

This section contains studies related to cryptocurrency price predictions. Some of these studies include the use of sentiment analysis, technical analysis, and tweet volumes as an indicator to predict certain cryptocurrencies. Moreover, several deep learning algorithms are explored.

Neural Network price prediction: The studies of Alonso-Monsalve et al. (2020), Critien (2021) and Li and Dai (2020) proposed several state-of-the-art deep learning algorithms such as CNN, LSTM, and even a combination of both, to predict cryptocurrency prices without the use of sentiment analysis.

Li and Dai (2020) implemented a hybrid CNN-LSTM model. The CNN section is in charge of the feature extraction and data input which comprises two convolutional layers as the pooling layers and a Dense layer. Moreover, the output of the CNN section is the input for the LSTM neural network. The LSTM neural network consists of one LSTM layer and a full connection layer. It resulted that the model achieved an F1-score of 0.69 to predict the direction of Bitcoin (Li and Dai, 2020).

Alonso-Monsalve et al. (2020) proposed to predict cryptocurrency using four different deep architectures such as the MLP, RBFNN, CNN, and a combination of CNN-LSTM networks. The authors examined six different cryptocurrencies for some neural networks, and concluded that the hybrid neural network (CNN-LSTM) outperformed the rest of the architectures with a result of a maximum accuracy of 0.61% (Alonso-Monsalve et al., 2020).

A study by Critien (2021) proposed to predict the price direction utilising three different neural networks such as LSTM, BiLSTM and CNN which were based on sentiment analysis features. Another approach in this study was to predict the closing price magnitude of the day. This study managed to achieve maximum accuracy of 64.2%

using the BiLSTM neural network to predict the price direction. A voting classifier was used to get the utmost result from the best-performed model for predicting the price direction and price magnitude, obtaining a maximum accuracy of 77.2%.

Classifiers for price predictions: A study by Valencia et al. (2019) proposed to use RF, SVM classifiers and a neural network known as MLP to predict the movements of four cryptocurrencies such as Bitcoin, Ethereum, Ripple and Litecoin. The RF achieved an accuracy of 61% predicting Bitcoin using only market data and the lowest on Ethereum with an accuracy of 28% using market data only.

Moreover, a study by Akyildirim et al. (2021) utilised RF classifier and other several machine learning algorithms in order to predict mid-price movements for Bitcoin future prices. The RF managed to achieve the highest average of an in-sample success ratio with 87%.

Twitter sentiments for price prediction: A study by Pant et al. (2018) implemented a RNN model with its variations such as GRU and LSTM in order to remove the vanishing gradient problem which RNNs suffer from. The Pearson's correlation coefficient was used to measure the relationship between the sentiments and price. The Pearson's value ranges from 1 and -1 where 1 means a strong relation and weak respectively. In this study, the negative sentiments and the fall in price had a coefficient of 0.41 whilst the positive sentiment and the rise in price had a value of 0.26. The price prediction accuracy for the RNN predictor was 77.62%.

Tweet volume for prediction: Abraham et al. (2018) proposed a cryptocurrency price prediction whereby utilising Twitter volume and Google Trends, as it resulted that the sentiments had a minor correlation since tweets were more neutral rather than objective. Therefore, the author decided to ignore it as it was not a reliable indicator. On the other hand, Google Trends was highly correlated with the price movement whether it was bullish or bearish. It managed to achieve a correlation coefficient of 0.817 with a p-value of 0.000. The p-value represents whether the correlation metrics are reliable; a smaller p-value suggests stronger evidence of an alternative hypothesis. In regards to the tweet volume, it achieved a higher correlation than the model using Google Trends with a coefficient of 0.841 and a p-value of 0.000.

Using other various neural networks: In the work proposed by Mittal et al. (2019), the author attempts to prove whether a correlation between Bitcoin price, Google search and Twitter exists by using various neural networks such as Linear Regression and Polynomial Regression. Polynomial Regression resulted that the overall accuracy for Google Trends and Tweet volume being 66.66% and 77.01% respectively. Using the Linear Regression neural network, the R2 Score and Pearson were taken to measure the correlation of Bitcoin price between tweet volume, Google Trends and Tweet sentiment. The R2 Score for Tweet Volume, Google Trends and Tweet Sentiments were 0.690, 0.755 and 0.49 for the latter. The Correlation Coefficient were 0.740, 0.790 and -0.300 respectively. Thus, having the sentiment as a not satisfactory result (Mittal et al., 2019). Moreover, the study by Wołk (2020) uses Google Trends data to predict six different cryptocurrencies named Bitcoin, Electroneum, Ethereum, Monero, Ripple and ZCash by implementing several neural networks such as AdaBoost, Support Vector Regression, Stochastic gradient descent, Gradient Boosting model, MLP, Decision tree, Bayesian Ridge Regression, ElasticNet and Hybrid networks.

2.2.6 Limitations of Related Work

The major limitations related to studies that were mentioned above include both textual and financial historical datasets (Kilimci, 2020; Pant et al., 2018), lack of detailed analysis (Wołk, 2020) and the quality of the content such as duplication of input by bots and advertisements (Valencia et al., 2019). Moreover, large data requires higher computational time and cost.

Consequently, the current study attempts to lessen these problems by depicting a more detailed result, mainly stemming from the preprocessing phases, by undertaking and testing the most adequate cleaning process for this social media content. This will ensure higher quality input data for the sentiment analyser to make it an efficient indicator.

Chapter 3. Research Methodology

This chapter discusses the research methodology of the prototype used to answer the research questions of the present study. Figure 3.1 demonstrates an overview of the phases that need to be undertaken to predict cryptocurrency price direction. These phases include data collection such as collecting Twitter posts and the prices of Bitcoin, cleaning and pre-processing of the data and the machine learning models which will be used for predicting the price direction and comparing them for the best model.

This design was considered since there were some limitations from the Twitter API regarding the limit of requests that one can have. An Influential user implementation using a personalised PageRank algorithm as used in some other studies (Cheuque Cerda and L. Reutter, 2019; Eliacik and Erdogan, 2018; Riquelme and González-Cantergiani, 2016; Wooley et al., 2019) was considered, however, it could not be implemented due to the limit of requests to process all the data since it would have taken a considerable amount of time.

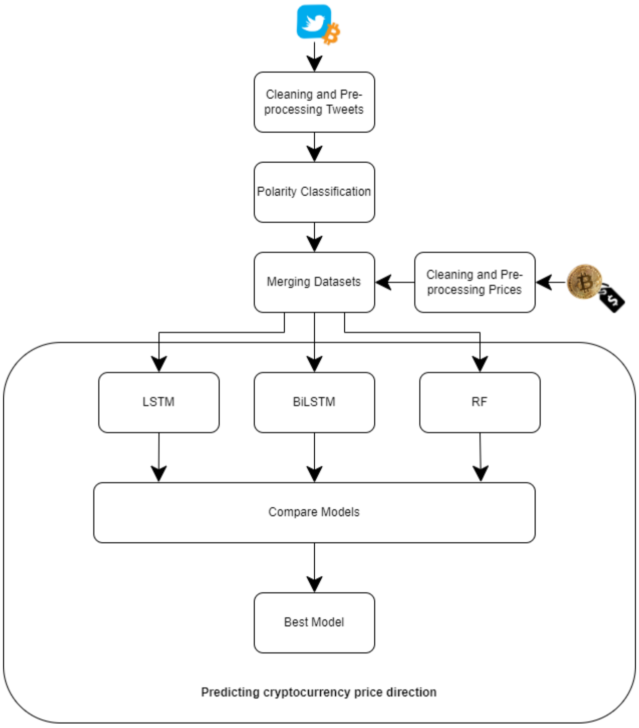
Cleaning and Pre-processing Tweets
Polarity Classification
Merging Datasets
Cleaning and Pre-processing Prices
LSTM
BiLSTM
RF
Compare Models
Best Model
Predicting cryptocurrency price direction

Figure 3.1: An overview of the whole process to predict cryptocurrency direction

3.1 Environment

The Python programming language was utilised and preferred as it become the most preferred language for data science due to its deep learning models, data handling and data analytics (Nagpal and Gabrani, 2019). The Python packages used throughout this research can be found in Table A.1.

3.2 Data Collection

One of the main pillars for prediction is data. This section discusses the data collection process for this research. The data required for this research include Bitcoin prices, Twitter posts and annotated Twitter posts.

3.2.1 Collecting Bitcoin prices

As mentioned in the previous chapter several sources exist to obtain Bitcoin prices at various intervals. Despite having a ready-made dataset from an openly shared repository like Kaggle, the CoinDesk API was the preferred choice since it provided the opportunity to choose between the needed date range for a daily interval (CoinDesk, 2021). Some common and important features of this dataset are the timestamp, the opening and closing price, and the highest and lowest price of every day.

3.2.2 Collecting Bitcoin tweets

In the previous related work section, Twitter is utilised to gather users' sentiment due to its popularity in this field. Comparably, to collect tweets related to Bitcoin, Kaggle is used. Kaggle is a source that provides a variety of datasets that can be used in research (Kaggle, 2022). The ready-made dataset that is used is named "Bitcoin tweets – 16M tweets". In the description part, it states that the date ranges from 1st of January 2016 until 29th March 2019 which contains the keywords 'bitcoin' and 'btc'. Some common features within this dataset were the identifier, timestamp, username and full name, the summation of retweets, likes and replies and the actual text of the tweet. Some of these common features were not required such as the likes, retweets and replies since the PageRank algorithm was discontinued due to the limitations of the Twitter API. This dataset was collected by Alexandre Boiullet using Tweepy and Twint which are software packages within the Python ecosystem to access Twitter's API.

3.2.3 Collecting Annotated Twitter Posts

A dataset containing the Twitter posts was examined to understand better which cleaning process is more appropriate for the chosen sentiment analysers on social media content since social media content contains an abundance of noise. This dataset is provided by the NLTK library which contains 5000 positive tweets and 5000 negative tweets similarly to the study by Elbagir and Yang (2018). This dataset is utilised to perform certain NLP tasks on social media content hence an experiment is conducted on this dataset to prove which cleaning process is more adequate for VADER and TextBlob Sentiment Analysers.

3.3 Data Exploration

The datasets containing Twitter posts were further analysed to identify the contents. Data exploration was used to confirm the size of the dataset, the quality of the textual content, the presence of tweets generated by bots and the detection of any null values within the dataset.

3.4 Data cleaning and pre-processing

As mentioned in the previous section, a dataset containing Twitter posts normally is in a state that contains an abundance of noise which will be a hindrance for the sentiment analysers to obtain the right sentiment, hence it is necessary to undertake certain cleaning phases as emphasized in the research by Pradha et al. (2019). The following cleaning combination steps were performed on the annotated dataset provided by the NLTK library (Bird et al., 2009).

Several combinations of cleaning processes were considered, such as setting capitalization, removing extra white spaces, tokenization, and lemmatization, fixing contractions, and removing stop words, numbers, hashtags and profile mentions.

In terms of capitalization, text was set to lower case while redundant URLs and Emoticons were eliminated from the tweets by utilising Regular Expressions (Regex) patterns within the Python's ecosystem similar to the study by Pant et al. (2018). The Python Contractions package was utilised to fix any contractions available within a tweet (Pascall, 2020). Contractions are words such as "don't" and "you're" which are split into two separate words such as "do not" and "you are".

Moreover, another important phase that took place in pre-processing the data was tokenization and lemmatization by utilising TweetTokenizer and WordNetLemmatizer respectively which are included in the NLTK library (Kraaijeveld and De Smedt, 2020). Once tweets were tokenized, the removal of hashtags, profile mentions, stop words and punctuation could be done much easier since words are in a form of arrays. The profile mentions were replaced with the word 'USER'. This follows a procedure similar to Critien (2021) and Pagolu et al. (2016). Concerning the hash sign ("#") a check-up occurred by using the Wordlist provided in NLTK which controlled whether the word after the hash sign exists in the dictionary if not it was dropped (Critien, 2021). Figure 3.2 depicts an overview of data cleaning and pre-processing steps for the dataset provided by the NLTK library.

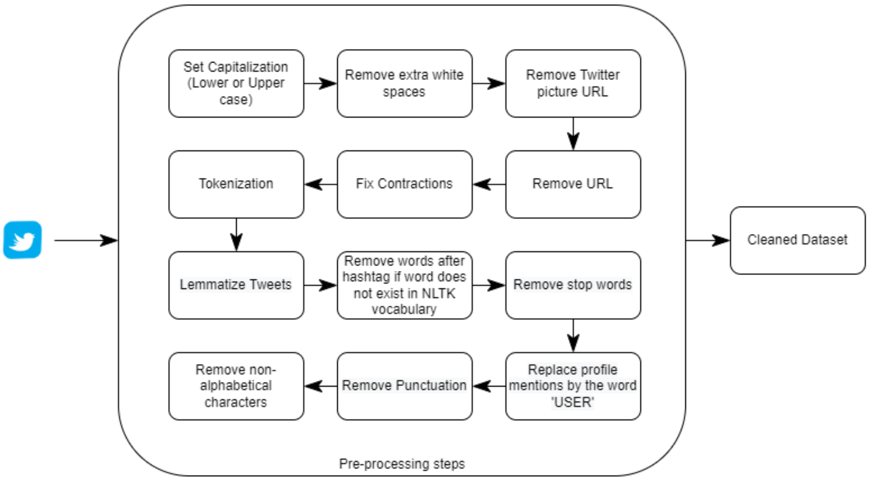
Set Capitalization (Lower or Upper case)
Remove extra white spaces
Remove Twitter picture URL
Tokenization
Fix Contractions
Remove URL
Lemmatize Tweets
Remove words after hashtag if word does not exist in NLTK vocabulary
Remove stop words
Remove non-alphabetical characters
Remove Punctuation
Replace profile mentions by the word 'USER'
Pre-processing steps
Cleaned Dataset

Figure 3.2: An overview of data cleaning and pre-processing steps

The best-performed cleaning process combination was utilised on the dataset containing Twitter posts related to Bitcoin in order to predict the price direction. The data exploration phase resulted in that several tweets were dated from 2009, hence, the first phase of the cleaning process was to trim all posts before 1st January 2017, since that was the initial date that data was making sense. Moreover, duplicated tweets were dropped, and non-English tweets were eliminated by utilising Python's package named "langdetect" as carried out in the study by Stelzmüller et al. (2021). Apart from eliminating duplicate tweets, the removal of bots was done by removing tweets containing the words "give away", "giving away", "join", "register", and "pump" as suggested in the study by Kraaijeveld and De Smedt (2020). Lastly, phrases with less than 4 words were eliminated similar to the study by (Critien, 2021). Figure 3.3 depicts the cleaning process performed on the dataset containing the Twitter posts related to Bitcoin. The phases performed are based on the results obtained from the dataset provided by the NLTK library and conducting data exploration.

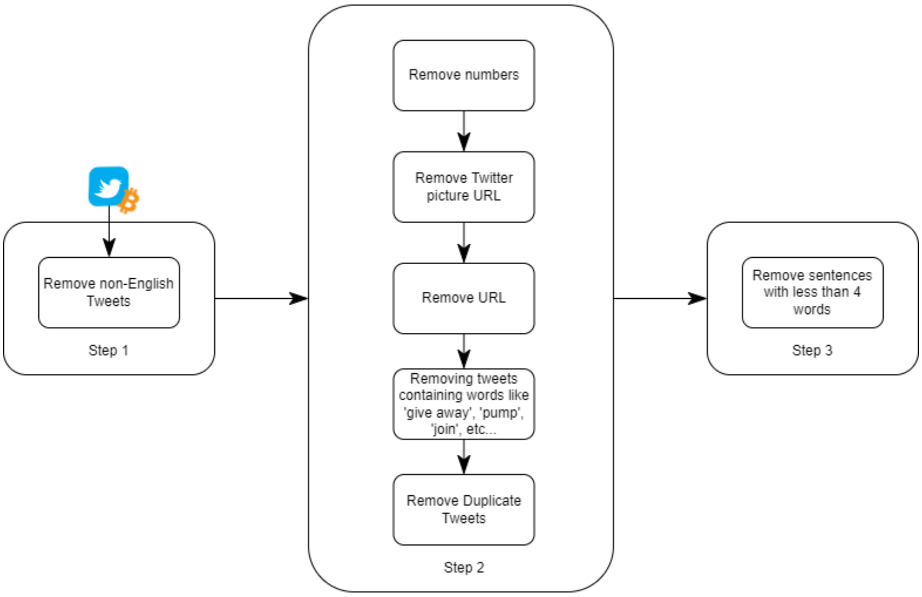
Remove non-English Tweets
Step 1
Remove numbers
Remove Twitter picture URL
Remove URL
Removing tweets containing words like 'give away', 'pump', 'join', etc...
Remove Duplicate Tweets
Step 2
Remove sentences with less than 4 words
Step 3

Figure 3.3: An overview of the cleaning processes used for tweets related to Bitcoin

Concerning the dataset containing the Bitcoin's prices which was collected from CoinDesk API, the field containing the timestamp was renamed and converted to *DateTime* data type to match the Bitcoin's tweets dataset, when merging occurs. Moreover, the High and Low fields were eliminated from the dataset, maintaining only the daily closing price.

3.5 Polarity classification

One of the research questions **(RQ2)** that this study attempts to answer is whether VADER outperforms TextBlob analyser in terms of extracting the polarity, therefore these two different sentiment analysers were tested on the dataset to understand which one of the two performed better on social media content based on the cleaning combinations depicted in Figure 3.2. The dataset provided by the NLTK library was used to understand which cleaning combination for the sentiment analysers performs better on social media content. The two sentiment analysers extracted polarity on the combination of pre-processing techniques that were mentioned in the previous section.

The best combination of cleaning processes on TextBlob and VADER sentiment analysers was chosen to eventually predict the price direction.

VADER and TextBlob were chosen as candidates since they are open source, attuned to social media content and used in previous studies (Abraham et al., 2018; Kraaijeveld and De Smedt, 2020; Mohapatra et al., 2019; Valencia et al., 2019). Figure 3.4 depicts an overview of the sentiment analysers and polarity classification process during the analysation of the annotated dataset.

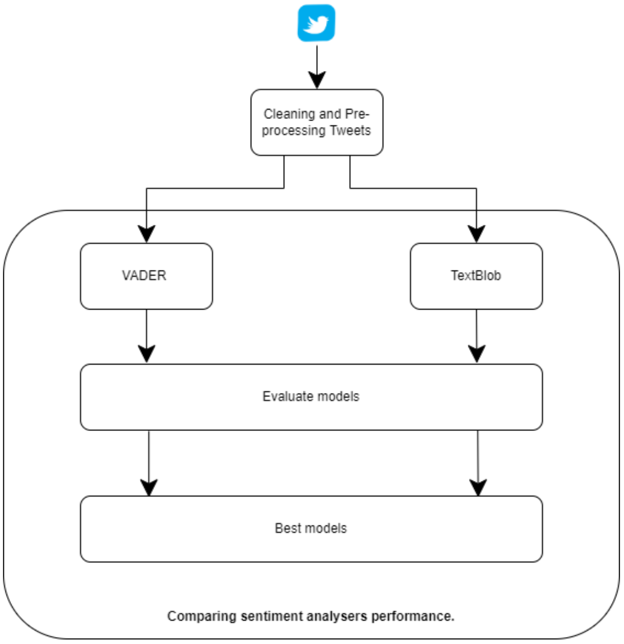
Cleaning and Pre-processing Tweets
VADER
TextBlob
Evaluate models
Best models
Comparing sentiment analysers performance.

Figure 3.4: An overview of the sentiment analysers and polarity classification

3.6 Grouping and Merging datasets

The main objective of this study is to predict the next day's change in direction of the closing price of Bitcoin, hence, the data obtained so far has to be grouped day by day for the models to perform daily predictions. Currently, the datasets containing the sentiments and prices are two separate datasets, hence, they need to be merged into one dataset. The financial dataset collected from CoinDesk API is in the daily interval, hence, there was no need of grouping it since it was already done. The process of grouping the dataset containing the tweets involves several phases. Primarily, using the mathematical *floor* function within the *DateTime* Python's package, the time from the timestamp attribute was set to zero to associate every tweet posted on the same day together. Secondly, the tweet volume was added as an additional feature as suggested by Abraham et al. (2018). This was obtained by counting the tweets when grouping per day. Lastly, polarity scores were averaged to get the average score of that particular day. Once the grouping process is completed, merging can be accomplished. The merging process was done on the timestamp that matches.

3.6.1 Features and Labels

The TextBlob features mentioned below are used to train the models:

- **Change Direction:** The target variable that this study is aiming to predict. Bitcoin's price changes direction of the day which can be 0 or 1 if the price goes down or up respectively.
- **Closing Price (USD):** Bitcoin's closing price of the day.
- **Polarity Score:** The sentiment score obtained from Twitter posts.
- **Subjectivity Score:** The subjectivity score obtained from Twitter posts.
- **Tweet volume:** The tweet volume in the daily interval, similar to Abraham et al. (2018).

The VADER features mentioned below are used to train the models:

- **Change Direction:** The target variable that this study is aiming to predict. Bitcoin's price change direction of the day which can be 0 or 1 if the price goes down or up respectively.
- **Closing Price (USD):** Bitcoin's closing price for a particular day.
- **Positive Score:** The average positive sentiment score obtained for that particular day.

- **Negative Score:** The average negative sentiment score obtained for that particular day.
- **Tweet volume:** The tweet volume in the record's interval, similar to Abraham et al. (2018).

3.7 Lagged Features

In time-series scenarios, forecasting problems must be re-framed as supervised learning problems (Mudassir et al., 2020). A method known as Multivariate Forecasting was implemented similarly in the study by Azaryan (2020) and Uras et al. (2020). A different number of input lag observations were considered, such as 1, 3 and 7 lagged time steps where data was shifted depending on how many lagged days were inputted. This data will be utilised for training the models. The number of output observations is 1 meaning that it will forecast the current time. This method converts the labelled features into var(t) for the original time steps and var(t-n) for nth-lagged features. Once reframed to supervised, the total features should accumulate depending on the number of lag observations inputted. Table 3.1 depicts an instance of Multivariate forecasting of 1-day lag feature, where reframed data such as var(t-n) is the nth-lagged feature that will be subjected to the models for training while the var(t) is without lag meaning the original time. Given the fact that the data has shifted to one day, subsequently, records containing Not a Number (NaN) value were discarded.

Date	**var1(t)**	**var2(t)**	**var1(t-1)**	**var2(t-1)**
D1	*S1*	*F1*	*NaN*	*NaN*
D2	*S2*	*F2*	*S1*	*F1*
D3	*S3*	*F3*	*S2*	*F2*
D4	*S4*	*F4*	*S3*	*F3*
D5	*S5*	*F5*	*S4*	*F4*

Table 3.1: Reframing time-series problem to supervised learning problem.

3.8 Neural Networks and Tuning configurations

In this research, **RQ1** addresses whether the next day's cryptocurrency price direction can be predicted by utilising social media sentiment analysis. Hence this problem can be classified as a binary classification problem since the prediction is based on the change in price whether it hikes or falls (Wooley et al., 2019). A classifier such as the RF classifier and two different neural network models such as LSTM and BiLSTM were used to answer the first research question which consequently answers **RQ3** and **RQ4**.

Different neural networks require different configurations to achieve the best performance possible. Finding the best configurations was challenging and time-consuming. These configurations include a different number of layers, neurons, and batch sizes. The LSTM and BiLSTM were configured using the Hyperbolic Tangent activation function as suggested by Farzad et al. (2019). The aforementioned activation function is for the input and hidden layers if available. The 'SoftMax' activation function was used for the output layer as it was utilised in multiple research (Jiang and Liang, 2017; Ly et al., 2018; Ulumuddin et al., 2020). A script was created for each neural network. Moreover, each model will iterate 5 times to get the mean accuracy of that model. Furthermore, each model was configured with different configurations in order to get the utmost performance on these machine learning models. All these combinations of different configurations are trained on a maximum of 10,000 epochs, yet an early stopping call-back function is utilised to eliminate overfitting. This function will stop the training if the validation loss will not be improved after 50 epochs.

Given the fact that exploring every possible scenario is impossible, a set of possible configurations for every model were assigned. Table 3.2 depicts the possible combinations of configurations values that are used to train the models which in total had accumulated to 225 different models.

Configurations	Lag Features	Layers	Neurons	Batch Sizes
Combinations Values	1, 3, 7	1, 2, 3	16, 32, 64, 128, 256	5, 20, 50, 80, 100

Table 3.2: BiLSTM and LSTM networks configurations.

3.9 Random Forest Classifier Hyper-Parameters

The combination of the RF tuning configuration was different compared to the neural networks since it included different parameters to configure the model. The hyper-parameters utilised for the RF apart from the lagged features were the number estimators and the criterion. Table 3.3 depicts the different combinations of parameters passed to configure the model. Certain hyper-parameters were selected as they were used similarly in the study by Felizardo et al. (2019). The number of estimators represents the number of trees in the forest. Moreover, criterions are the function to measure the quality of the split. The max depths were not considered as it might cause the model to overfit.

Configurations	Lag Features	Estimators	Criterions
Combinations Values	1, 3, 7	5, 50, 100, 250, 500	Gini, Entropy

Table 3.3: Random Forest classifier configurations

3.10 Further price predictions

Based on the results obtained and based on the hypothesis that an up-trend will trigger high sentiment volume, the machine learning models were executed for the last time on the same dataset yet with a smaller range of dates which ranges from 1st August 2018 to 23rd November 2019. The reason being since there was a higher correlation between the target variable and the predictor variables hence it might enhance the performance of the models. This period was chosen because it is well known to be a period in which a huge up-trend was observed.

3.11 Data Evaluation

Using the seaborn library, a heatmap was used to display the correlation of all the features. The Pearson R correlation coefficient was utilised to measure the correlation between the features that will be used as predictors to predict the change direction of the price. Moreover, this method was used since it returns an object containing the correlation coefficient, p-value as was used in the research by Abraham et al. (2018). The Precision Score, Recall and F1-Score for the best models were implemented to understand the reliability of the performance achieved. These metrics were used as suggested by Critien (2021) and Pant et al. (2018).

Chapter 4. Analysis of Results and Discussion

This chapter provides an analysis of the results obtained from the approach described in the research methodology. This section includes the data analysis of the sentiment analysers, RF Classifier, neural network models with and without different lagged features and different tuning configurations of the models which will lead to the best performing model and be able to answer the proposed research questions.

4.1 Data Exploration

The data exploration phase acquired a better understanding of the three datasets. The title and description of the dataset containing the Twitter posts related to Bitcoin were misleading since the dataset contained over 20 million tweets and certain tweets were dated from 2009 till the 23rd of November 2019. Furthermore, the feature containing the user opinion contained an abundance of noise and the presence of tweets generated by bots which were removed in the cleaning process. Concerning the financial dataset, Figure 4.1 depicts the time-series chart of the financial dataset.

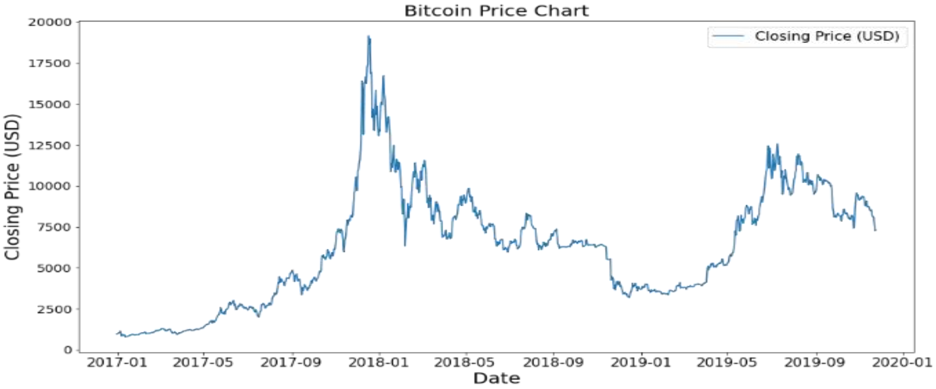

Figure 4.1: An overview of the Bitcoin Price chart between 01-01-2017 and 23-11-2019

4.2 TextBlob and VADER sentiment analysers

As discussed in the previous chapter the VADER and TextBlob sentiment analysers were examined on a dataset containing 5000 positive and 5000 negative tweets provided by the NLTK library. Both sentiment analysers were subjected to a combination of cleaning processes to determine which cleaning process achieved the best accuracy possible on a particular sentiment analyser for polarity extraction on social media content. It resulted that the best cleaning combination for both sentiment analysers was removing the URLs while the worst one was when all the cleaning combinations were included. Figure 4.2 and Figure 4.3 depict the confusion matrices of the best and worst cleaning combinations for the TextBlob and VADER sentiment analysers respectively.

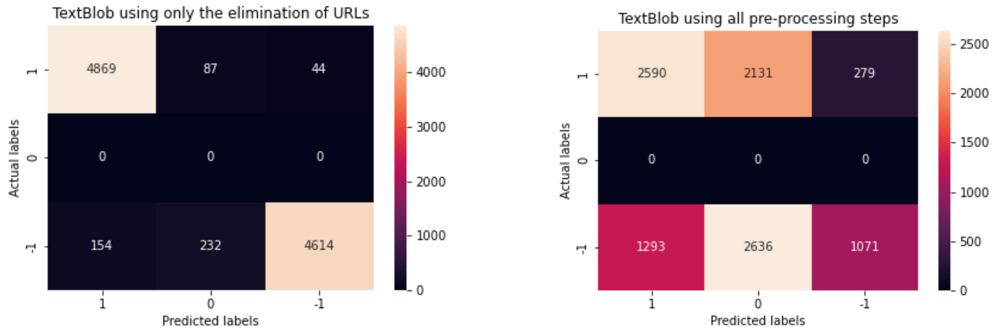

Figure 4.2: TextBlob Confusion Matrices

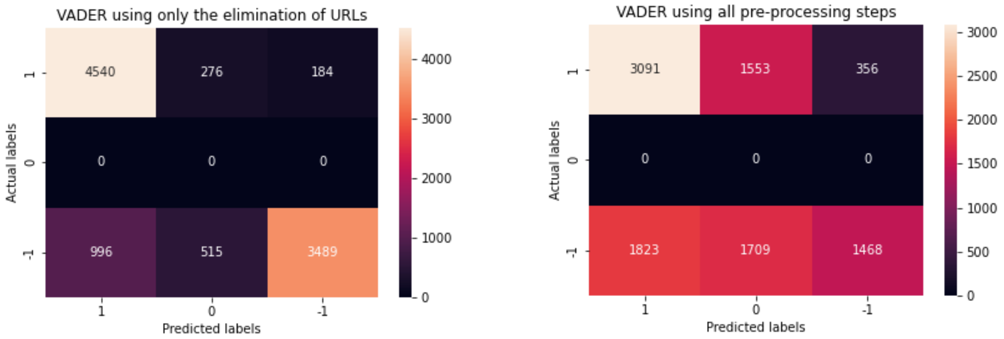

Figure 4.3: VADER Confusion Matrices

One can notice that the confusion matrices depict that on the dataset provided by the NLTK library, the TextBlob analyser achieved the best performance. Table 4.1 and Table 4.2 display the results achieved for both sentiment analysers on the full combination of pre-processing steps mentioned (Full Cleaned Tweets) and eliminating only the URLs data (Cleaned Tweets URLs) which resulted to be the best cleaning processes on social media content.

VADER	Accuracy %	Precision %	Recall %	F1-Score %
Full Cleaned Tweets	45.59	47.79	30.39	35.13
Cleaned Tweets (URLs)	80.29	59	53.53	55.55

Table 4.1: The results obtained from the VADER sentiment analyser

TextBlob	Accuracy %	Precision %	Recall %	F1-Score %
Full Cleaned Tweets	36.61	48.68	24.41	30.68
Cleaned Tweets (URLs)	94.83	65.33	63.22	64.23

Table 4.2: The results obtained from the Text Blob sentiment analyser

The VADER sentiment analyser managed to achieve an accuracy of 80.29% and an F1-Score of 55.55%, comparably, the TextBlob sentiment analyser achieved an accuracy of 94.83% with an F1-Score of 64.23% setting this model more reliable than the other. The result obtained from this experiment demonstrates that these particular sentiment analysers perform better on a particular cleaning process. This is so since these sentiment analysers are built on purpose to extract polarity from social media content. For instance, emoticons provide sentiment for these particular analysers since they are sensitive to intensity, therefore, including all the aforementioned pre-processing steps would remove such important components from the Twitter posts. On the other hand, other different sentiment analysers which were not attuned for social media content would have performed differently. These results conclude that on the dataset provided by the NLTK library, the TextBlob sentiment analyser outperforms VADER which gives a clear answer to **RQ2**.

4.3 Sentiment Analysis

Concerning the dataset containing the Bitcoin tweets, despite the results achieved between the two sentiment analysers in the previous section, further pre-processing must be performed to remove non-English tweets and tweets generated by bots such as duplicated tweets, and tweets containing certain words that may lead to an auto-generated tweet, that has deducted the number of tweets to over 6 million. Figure 4.4 and Figure 4.5 illustrate the distribution of polarity extracted from both analysers before and after removing neutral tweets. The polarity extracted tends to be more positive regardless of the price similarly to the study by Abraham et al. (2018). In fact, after eliminating neutral tweets 78.4% were positive for TextBlob and 71.1% for VADER while negative sentiment was low compared to the positive sentiment. This demonstrates that people tend to post more positively on Bitcoin.

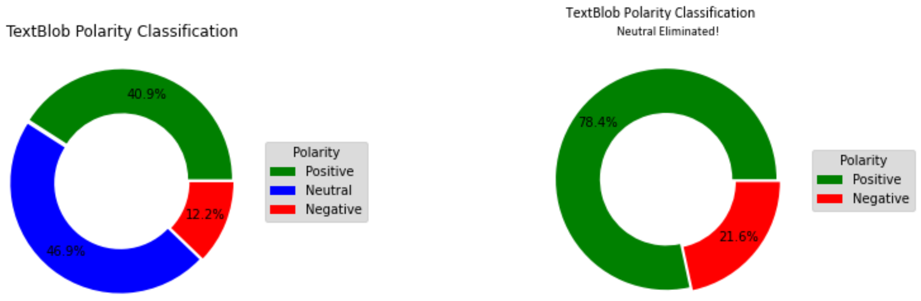

Figure 4.4: TextBlob polarity classification with and without neutral tweets

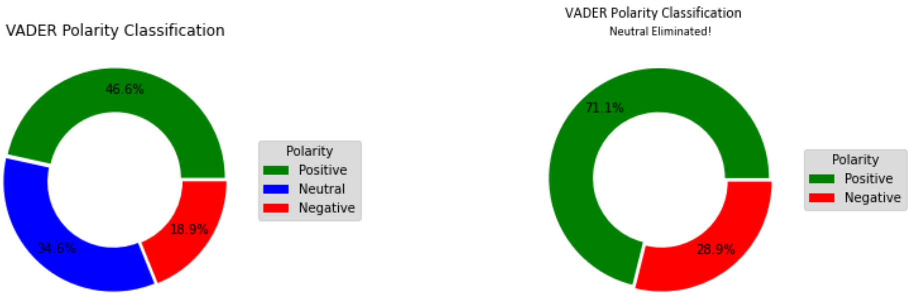

Figure 4.5: VADER polarity classification with and without neutral tweets

Moreover, the dataset containing the sentiments extracted was merged with the financial dataset. Figure 4.6 depicts a correlation heatmap of the features using TextBlob and VADER sentiment analysers. The results obtained from the Pearson R correlation coefficient demonstrated a low correlation between the features. The highest correlation is between the closing price and the tweet volume with a coefficient of 0.43. This demonstrates that the sentiment analysis is not an efficient indicator to predict the price direction on this particular dataset. The features related to the VADER analyser are slightly more correlated than the TextBlob analyser.

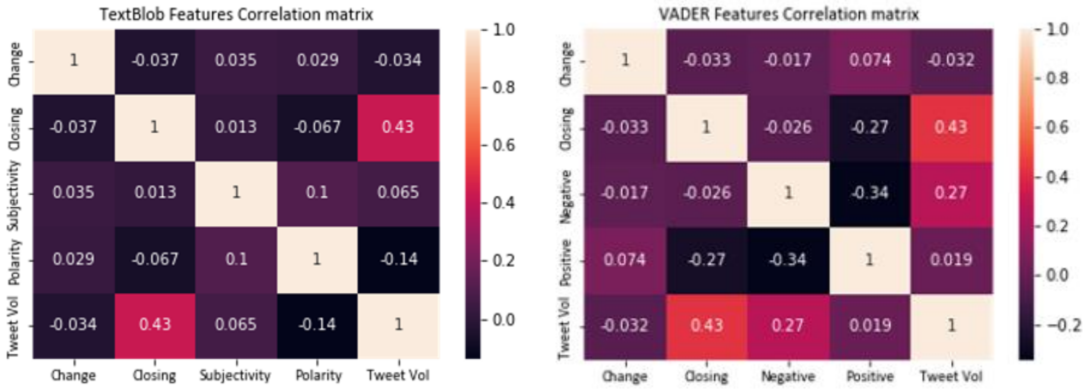

Figure 4.6: Heatmap correlation matrix of all the features.

Figure 4.7 and Figure 4.8 depict a correlation time-series line chart to understand when the data is more correlated. The time-series correlation of the features was to a certain extent, due to the size of the dataset since the dates range from 1st January 2017 to the

23rd of November 2019. Both figures demonstrate that the data resulted to be more correlated from mid-2018 onwards.

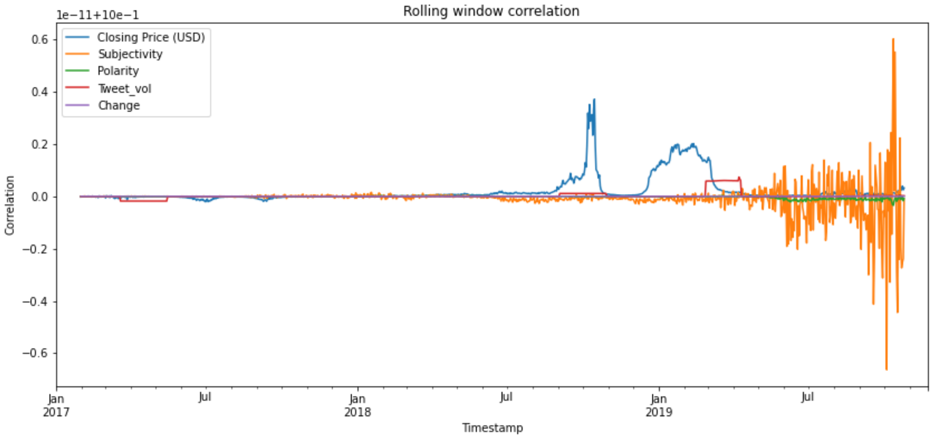

Figure 4.7: TextBlob time-series rolling window correlation.

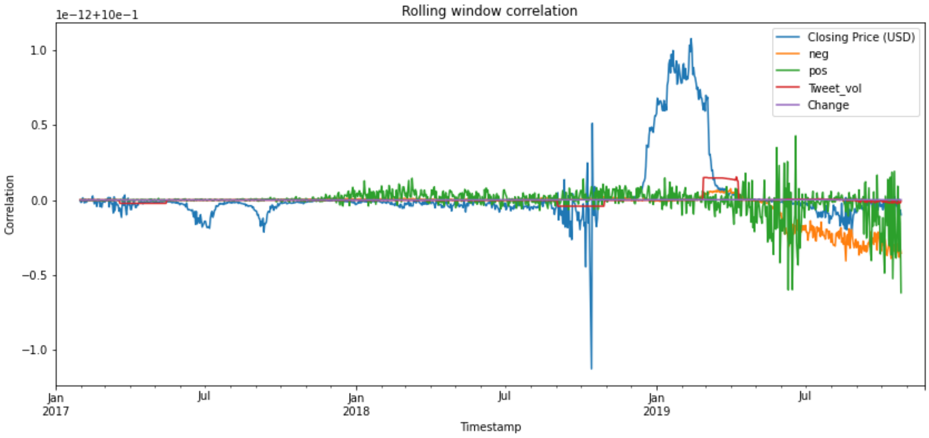

Figure 4.8: VADER time-series rolling window correlation.

Figure 4.9 illustrates the accumulation of the sentiment on the day. In fact, in the year 2019 the sentiment score was making more sense since every day there was a considerable amount of Twitter posts towards Bitcoin compared to 2017 and 2018. This demonstrates that Bitcoin was getting more popular after the bull run at the end of 2018 and therefore people started to tweet even more. Given the fact that the price of Bitcoin in 2019 was approaching the end of a bear market user sentiment was changing.

Subsequently, users start to tweet positively to influence other users within the cryptocurrency market. One can notice that the positive sentiment volume was always higher than the negative one. Positive fundamentals are expected to trigger an increase in the volume of positive sentiment and therefore the market value will increase. Effectively an increase in volume, and the market value will trigger more sentiment and influence other users to enter the cryptocurrency market.

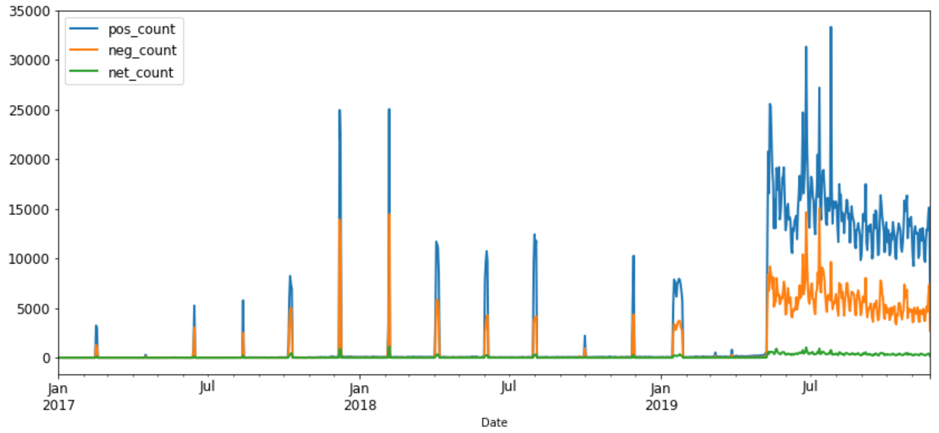

Figure 4.9: Sentiment score counted by day

4.4 Predicting cryptocurrency price direction without lagged features

Given the fact that the price is highly volatile, predicting the price direction was not an easy task. The results obtained from the machine learning models predicted the price direction to a certain extent. Figure 4.10 depicts the best results obtained from the trained models without introducing any lagged features. The mean and maximum accuracy is based on 5 iterations for each model. The best-performed model based on the VADER's sentiment features was the BiLSTM with an overall accuracy of 60.47% and a maximum accuracy of 60.85% and an F1-Score of 55.01% while the worst performance was done by the RF classifier with a mean accuracy of 52.45% and a maximum accuracy of 53.77% and an F1-Score of 52.30%. Using TextBlob features the models achieved a lower performance compared to VADER with the LSTM as the best model with an overall accuracy of 56.22% and a maximum accuracy of 58.01% while the worst performance was done by the RF classifier with an overall accuracy of 47.73%, a maximum accuracy of 50.47% and an F1-Score of 46.97%. In relation to **RQ3,** the sentiment analysis tended to be efficient to a certain extent nonetheless since there was a low correlation the performance was limited. Furthermore, to improve the performance of the models several lagged features were introduced.

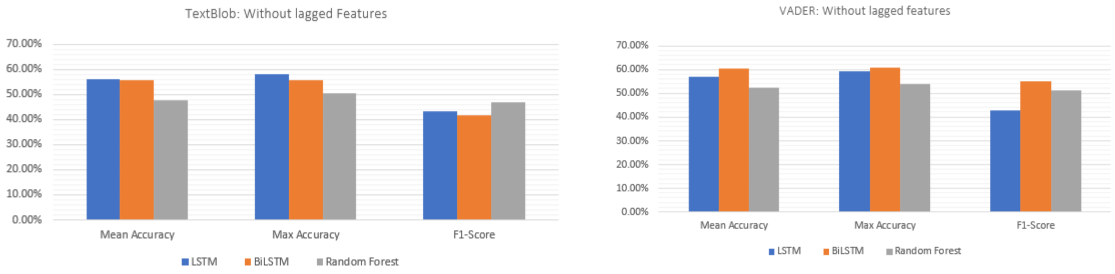

Figure 4.10: Models performance without lagged features

4.5 Predicting cryptocurrency price direction with lagged features

Introducing lagged features has helped the models to enhance their performance. Figure 4.11 depicts the mean, maximum accuracies, and F1-Score for the trained models based on the TextBlob features and different temporal lagged features. One can notice that the LSTM achieved the highest maximum accuracy of 62.09% to predict the next day's price direction. Moreover, the LSTMs performed better than the RF Classifier with a 1-day and 3-day lagged feature, yet the LSTM achieved the highest accuracy with a 3-day lagged feature. These results demonstrate that the best-performed model to predict the direction change of Bitcoin using lagged features is the LSTM since it achieved the highest maximum accuracy. On the other hand, the BiLSTM achieved more consistent results since the variation was low. This concludes that to predict the price direction, introducing lagged features demonstrated that the model performed better in particular for the 3-day lagged features. Subsequently, this will reflect on processing power since more data is inputted into the models for training, hence, more time-consuming.

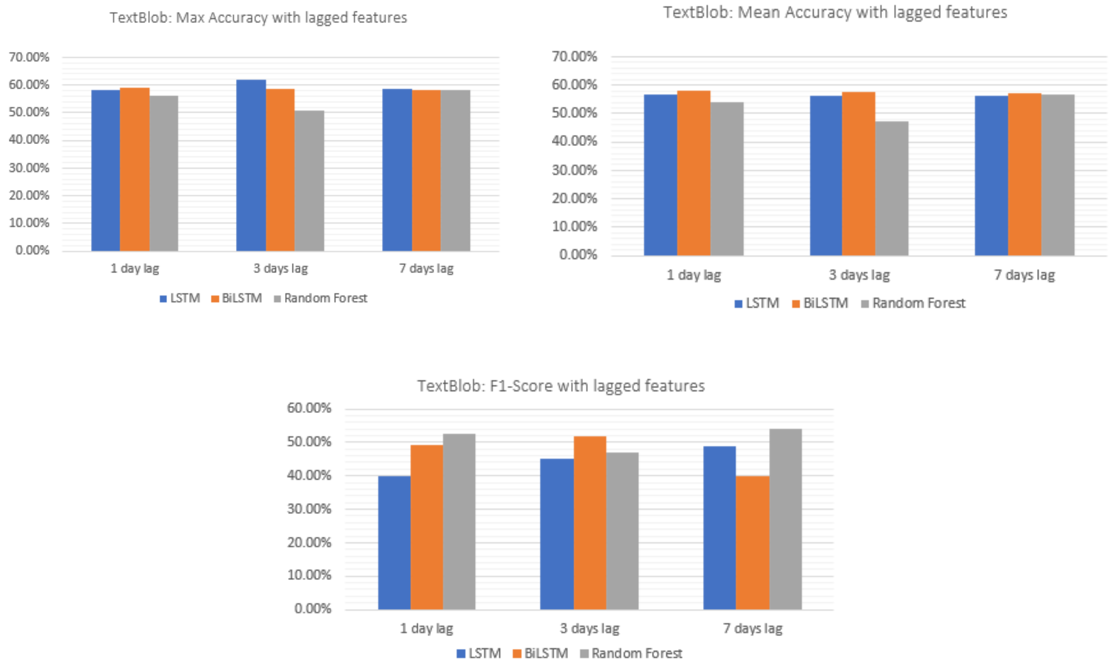

Figure 4.11: Mean, Maximum accuracy and F1-Score obtained from the models over different TextBlob lagged features.

Concerning the VADER lagged features, Figure 4.12 depicts the best results obtained from the trained models. The best-performed model was the LSTM using 7-day lagged features and manage to achieve maximum accuracy of 60.48%, an overall accuracy of 56.47% and an F1-Score of 54.63%. In terms of lag, the best-performed was on the 3-day lagged features. The result obtained did not outperform the one obtained by the TextBlob sentiment analyser features.

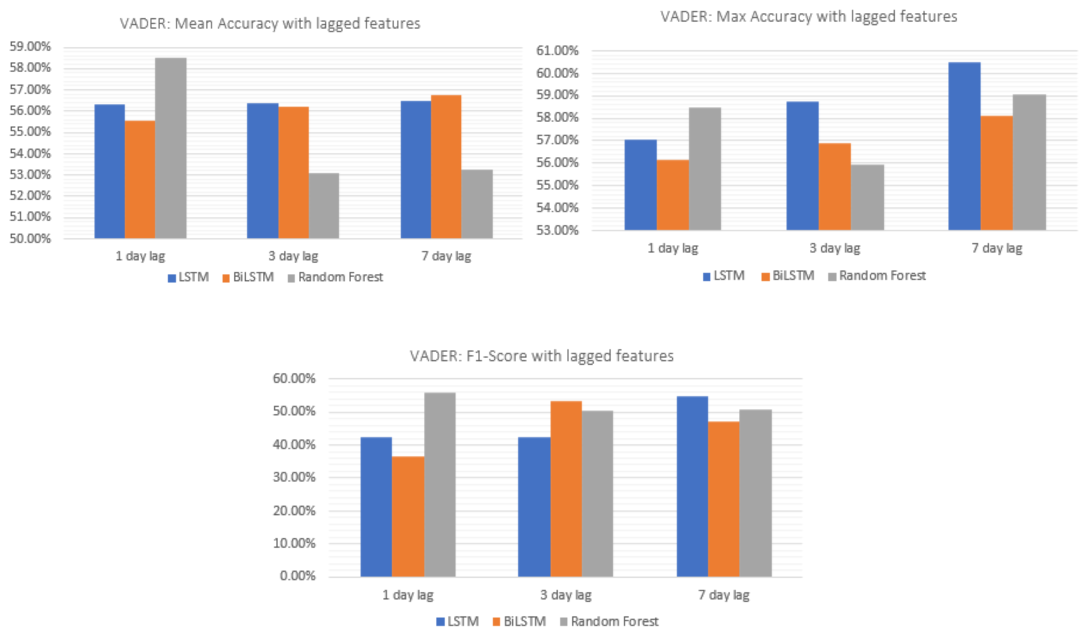

Figure 4.12: Mean, Maximum accuracy and F1-Score obtained from the models over different VADER lagged features.

4.6 Further price direction predictions

As mentioned in the previous chapter, predictions were further restricted to a smaller date range. This is so since a higher correlation was depicted in Figure 4.7 and Figure 4.8. Figure 4.13 depicts the correlation heatmap of the dataset with a smaller range of dates, hence, a higher correlation for both sentiment analysers. The Tweet volume tends to be more efficient than sentiment since there was a higher correlation.

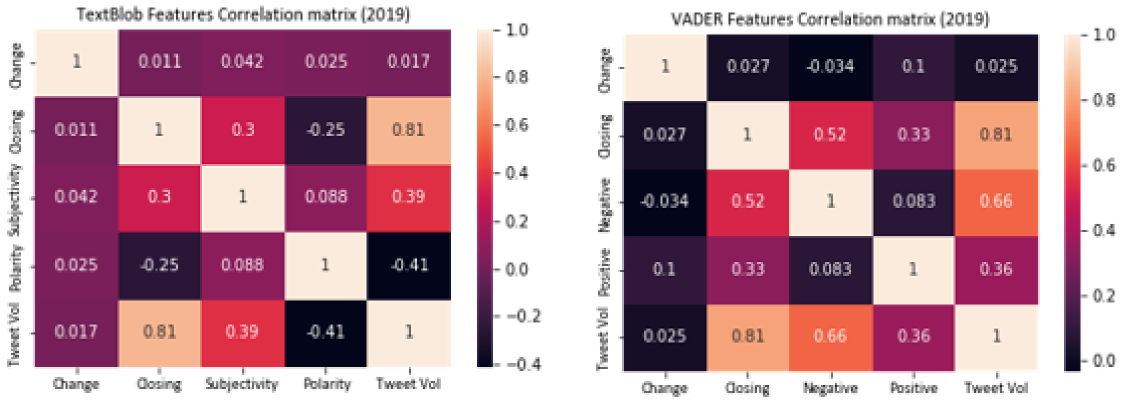

Figure 4.13: Heatmap correlation matrix with higher correlated features

Utilizing TextBlob analyser without lagged features, the LSTM managed to achieve an overall accuracy of 63.54% and a maximum accuracy of 64.58% with an F1-Score of 63.47%. Comparably, the BiLSTM achieved an overall accuracy of 61.88%, a maximum accuracy of 63.54% and an F1-Score of 61.50%. The RF classifier achieved an overall accuracy of 58.75%, a maximum accuracy of 62.50% and an F1-Score of 57.56%.

In regards to the VADER sentiment analyser, the LSTM achieved an overall accuracy of 56.88%, a maximum accuracy of 57.29% and an F1-Score of 52.80%. The BiLSTM achieved an overall accuracy of 55.21%, a maximum accuracy of 56.25% and an F1-Score of 55.20%. The RF classifier manages to achieve an overall accuracy of 51.46%, a maximum accuracy of 59.38% and an F1-Score of 51.13%. Figure 4.14 depicts the results of the best-performed models on both sentiment analysers without introducing any lag.

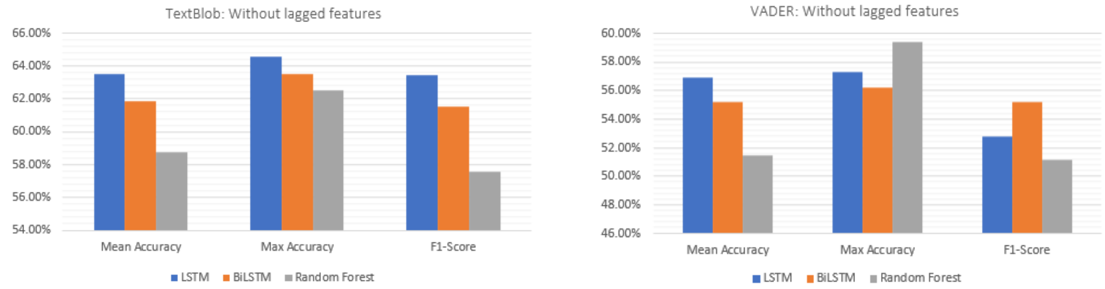

Figure 4.14: Models performance without lagged features on a higher correlated dataset

In relation to the models trained on TextBlob lagged features, the LSTM managed to achieve the best performance on a 1-day lag with an overall accuracy of 62.92%, a maximum accuracy of 63.54% and an F1-Score of 62.05% while the BiLSTM achieved the best performance using 7-day lag features with an overall accuracy of 56%, a maximum accuracy of 60%, with an F1-Score of 55.88%. The third model, the RF classifier, obtained the performance on the 7-day lagged features with an overall accuracy of 56%, a maximum accuracy of 58.95% with an F1-Score of 55.13%. Figure 4.15 depicts the performance of TextBlob with different lagged features.

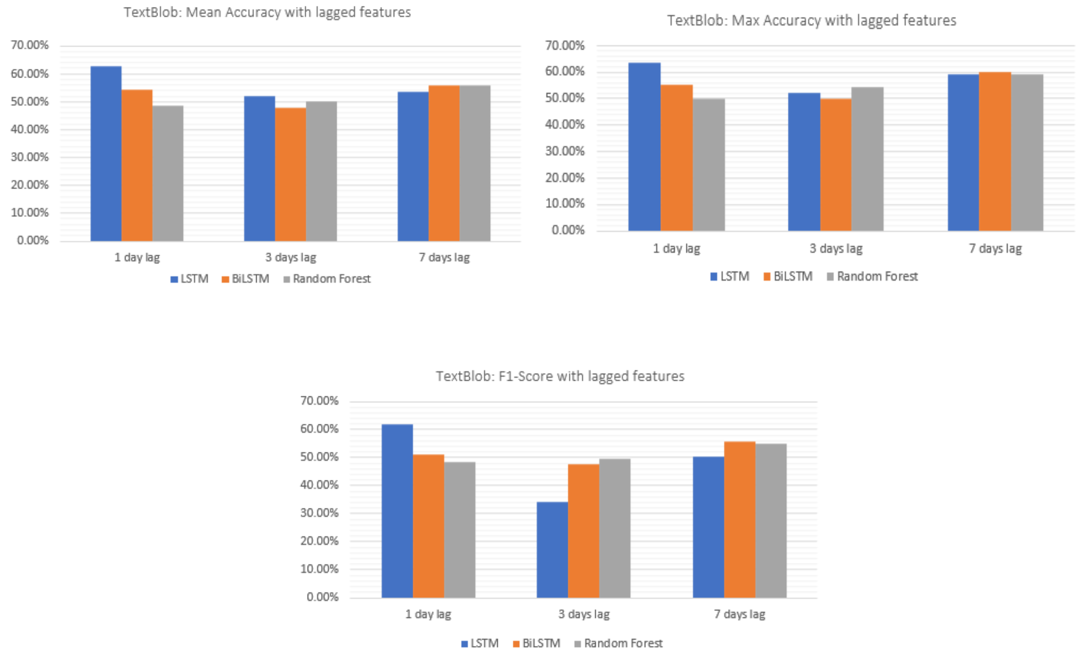

Figure 4.15: Mean, maximum accuracy and F1-Score of the models with TextBlob lagged features.

Figure 4.16 illustrates the best-performed models based on VADER's different lagged features. The best performance of the LSTM was on a 3-day lag with an overall accuracy of 60.63%, a maximum accuracy of 62.50% with an F1-Score of 60.41%. The BiLSTM best-performed model was on the 7-day lag with an overall accuracy of 59.79%, a maximum accuracy of 62.11% and an F1-Score of 59.10%. The RF classifier achieved the best performance using 7-day lag features with an overall accuracy of 55.16%, a maximum accuracy of 57.90% with an F1-Score of 54.36%. Each model performed better compared to the results obtained on the entire dataset. This is so since there was a better correlation between the target variable and the predictor variables and less chance for classification error.

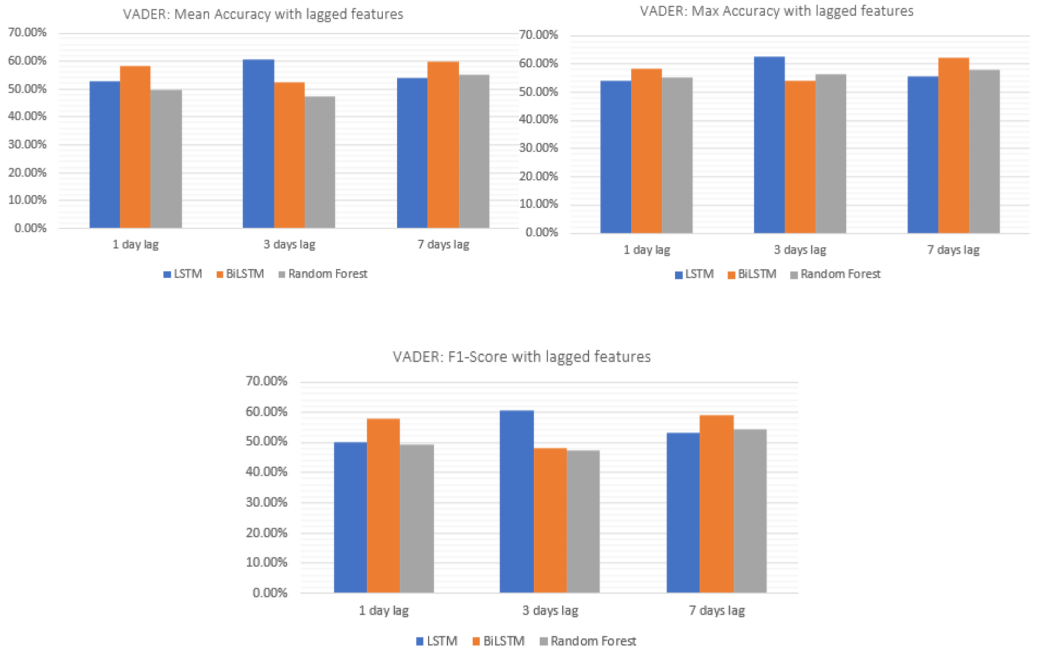

Figure 4.16: Mean, maximum accuracies and F1-Scores of the models using VADER lagged features.

Table 4.3 demonstrates the best tuning configuration for neural network models and the RF classifier with or without lagged features to predict the price direction of Bitcoin cryptocurrency utilizing sentiment analysis. These results will conclude to answer the research questions of this study.

	BiLSTM	LSTM	RF Classifier
Sentiment Analyser	TextBlob	TextBlob	TextBlob
Lagged Features	0	0	0
Layers	2	2	
Neurons	64	256	
Batch Size	5	20	
Input Activation Function	Tanh	Tanh	
Output Activation Function	SoftMax	SoftMax	
Loss Function	CC	CC	
Early Stopping Parameter	Validation Loss	Validation Loss	
Early Stopping Patience	50	50	
Maximum Accuracy	63.54%	64.58%	
Mean Accuracy	61.88%	63.54%	
Mean F1-Score	61.50%	63.47%	
Estimator			5
Criterion			Entropy
Maximum Accuracy			62.5%
Mean Accuracy			58.75%
Mean F1-Score			57.56%

Table 4.3: Best performed models' configuration

Concerning **RQ1,** Table 4.3 depicts that the LSTM has a maximum accuracy of 64.58% confirming that the next day's price direction can be predicted to a certain extent with this particular data. Moreover, the results obtained were reported to be better than the study by Alonso-Monsalve et al. (2020). In addition, the best-performed model was better than the Direction-BiLSTM model conducted by Critien (2021). On the other hand, the results did not manage to outperform the overall result in the study by Critien (2021). However, it should be noted that the present study did not implement the prediction of the price magnitude and a voting classifier.

Moreover, with **RQ2** VADER did not outperform the TextBlob analyser in extracting polarity in social media content. In addition, the highest accuracy was obtained by the LSTM resulting in the best-performed model in the present study which was based on TextBlob lagged features including the tweet volumes. However, one should note that the VADER features resulted to be slightly more correlated than the one of TextBlob. Certain studies use pre-processing steps without analysing the data to understand the ideal steps to perform. In this case, since these sentiment analysers are attuned to social media content minimal cleaning processes were required.

Furthermore, to answer **RQ3** social media sentiment analysis has proven that it is not an efficient indicator since there was a low correlation toward the target variable. Moreover, restricting the dataset, the VADER features had a coefficient of 0.66 for the negative sentiment and 0.36 for the positive sentiment. With the TextBlob features, subjectivity had a coefficient of 0.39 while the polarity had a coefficient of -0.41. This resulted an improvement compared to the correlation achieved on the entire dataset. The Tweet Volume obtained a higher correlation with a coefficient of 0.81 towards the closing price making it a better indicator. Nonetheless, the user's opinion tends to be positive regardless of whether the price is going down. Similar scenarios occurred in the study by Abraham et al. (2018), and Wołk (2020).

With **RQ4,** the LSTM provided the best performance with a maximum accuracy of 64.58%, placing it as the most preferred model to predict the price direction of the asset using social media sentiment analysis. The BiLSTM with a maximum accuracy of 63.54% and the RF classifier had the least maximum accuracy of 62.5%. With such results, one can conclude that the LSTM in the present study achieved higher accuracy in predicting the price direction of the Bitcoin cryptocurrency than in the study by Valencia et al. (2019).

Chapter 5. Conclusion and Recommendations

This study attempted to prove that the price direction of cryptocurrency, in particular Bitcoin, can be predicted using social media sentiment analysis from a social platform such as Twitter. Given the fact that NLP makes it possible to measure polarity, it was worth trying to use sentiment analysis to predict Bitcoin movements. This was approached by using two different sentiment analysers attuned to social media content such as the TextBlob and VADER sentiment analyser to extract users' opinions. Several combinations of pre-processing steps were considered to understand which cleaning process is preferred for these particular sentiment analysers. Moreover, once polarity was classified, features were inputted into the deep learning models such as LSTM and BiLSTM to predict the next day's change direction of the price. Additionally, an ensemble learning model such as RF classifier was also considered to predict the next day's price direction. In addition, 1, 3, and 7-day lagged features were introduced to enhance the model's performance as previous observations or features can influence or contain correlation towards the change direction. To conclude, based on the correlation metrics, another execution of the models was conducted to enhance the model since data had a higher correlation.

Every investor would like a model that predicts the direction of an asset without any failures, but that is not always the case since many variables need to be considered. This research managed to predict the next day's price direction to a certain extent. The maximum accuracy achieved to predict price direction was 62.09% using TextBlob sentiment analyser for polarity extraction with 3-day lag features and LSTM as a deep learning model. It is worth mentioning that the dataset was fairly large ranging from 1st January 2017 to 23rd November 2019, thus, this had affected terms of correlation and classification error. By reducing the dataset where features had a higher correlation the

LSTM model manage to achieve maximum accuracy of 64.58%. Moreover, the Bitcoin price movement has its peaks and lows, but Twitter sentiment tended to be more positive regardless of the market trend, whether is bullish or bearish. Moreover, according to Hanna et al. (2020) when the fundamentals are positive there will be more positive sentiment therefore the market value will increase. Effectively an increase in the market value will trigger more sentiment and influence other users to enter the cryptocurrency market.

5.1 Limitations

A limitation that occurred throughout this research was the Twitter API limit of requests since it offers 500 requests every 15 minutes. Moreover, to collect the followers of a user, 2 requests were needed, one to obtain the user object and another to get the followers' count. This would have required 37.5 days only to collect the followers of the unique users within the dataset since there were around 900,000 unique users. Time was limited therefore this approach was discontinued. Consequently, this has limited the implementation of the PageRank algorithm to distribute weight according to influence ranking.

5.2 Recommendations for Future Work

Further studies should consider introducing weights to sentiments based on the influence potential of a person or tweet. This can be done by implementing a Personalized PageRank algorithm which enhances the model. It will measure the weight of a particular sentiment based on the number of followers of a user and retweets, likes, and replies to a tweet. This will assist the model as each positive or negative sentiment is not identical as one can have more influence than the other. Another approach for distributing sentiment weights can be done similarly in the study by Mohapatra et al.

(2019) where the compound score is subjected to an equation to calculate the score based on followers, retweets and likes.

Another recommendation to enhance the model is to collect more data from different sources such as the Reddit platform or even financial news such as the Federal Reserve since fundamental analysis affects the global economy, hence, investors tend to cash out from the markets. The same approach can be implemented on other cryptocurrency assets. Additionally, one can consider using other different deep learning models and ensemble learning methods to predict the direction of an asset. It is worth mentioning that having more data will result in a more accurate model but more time-consuming to process.

To conclude this research has proven that the change in direction of the Bitcoin cryptocurrency can be predicted or verified to a certain extent and with the right cleaning process, introducing sentiment weights and lagged features might improve the model and possibly even make a profit.

Bibliography

ABD AZIZ, A. S., NOOR, N. A. M. & AL MASHHOUR, O. F. 2022. The Money of The Future: A Study of The Legal Challenges Facing Cryptocurrencies. *BiLD Law Journal,* 7**,** 21-33.

ABRAHAM, J., HIGDON, D., NELSON, J. & IBARRA, J. 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review,* 1**,** 1.

AKYILDIRIM, E., CEPNI, O., CORBET, S. & UDDIN, G. S. 2021. Forecasting mid-price movement of Bitcoin futures using machine learning. *Annals of Operations Research***,** 1-32.

ALONSO-MONSALVE, S., SUÁREZ-CETRULO, A. L., CERVANTES, A. & QUINTANA, D. 2020. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. *Expert Systems with Applications,* 149**,** 113250.

ALOTHALI, E., ZAKI, N., MOHAMED, E. A. & ALASHWAL, H. Detecting social bots on twitter: a literature review. 2018 International conference on innovations in information technology (IIT), 2018. IEEE, 175-180.

AZARYAN, A. 2020. Machine learning approaches for financial time series forecasting.

BALFAGIH, A. M. & KESELJ, V. Evaluating sentiment c1assifiers for bitcoin tweets in price prediction task. 2019 IEEE International Conference on Big Data (Big Data), 2019. IEEE, 5499-5506.

BAUR, D. G. & DIMPFL, T. 2021. The volatility of Bitcoin and its role as a medium of exchange and a store of value. *Empirical Economics,* 61**,** 2663-2683.

BIRD, S., KLEIN, E. & LOPER, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*, "O'Reilly Media, Inc.".

BONTA, V., KUMARESH, N. & JANARDHAN, N. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science Technology,* 8**,** 1-6.

CHEUQUE CERDA, G. & L. REUTTER, J. Bitcoin price prediction through opinion mining. Companion Proceedings of The 2019 World Wide Web Conference, 2019. 755-762.

COINDESK. 2021. *CoinDesk* [Online]. Available: https://www.coindesk.com/price/bitcoin/ [Accessed 6th August 2021].

COINMARKETCAP. 2021. *CoinMarketCap,* [Online]. Available: https://coinmarketcap.com/currencies/bitcoin/ [Accessed November 4, 2021].

CRITIEN 2021. Sentiment Analysis to Predict Cryptocurrency Prices.

DEVLIN, J., CHANG, M.-W., LEE, K. & TOUTANOVA, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

ELBAGIR, S. & YANG, J. Sentiment analysis of twitter data using machine learning techniques and scikit-learn. Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, 2018. 1-5.

ELIACIK, A. B. & ERDOGAN, N. 2018. Influential user weighted sentiment analysis on topic based microblogging community. *Expert Systems with Applications,* 92**,** 403-418.

FARZAD, A., MASHAYEKHI, H. & HASSANPOUR, H. 2019. A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Computing Applications,* 31**,** 2507-2521.

FELIZARDO, L., OLIVEIRA, R., DEL-MORAL-HERNANDEZ, E. & COZMAN, F. Comparative study of bitcoin price prediction using WaveNets, recurrent neural networks and other machine learning methods. 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 2019. IEEE, 1-6.

FERDIANSYAH, F., OTHMAN, S. H., RADZI, R. Z. R. M., STIAWAN, D., SAZAKI, Y. & EPENDI, U. A lstm-method for bitcoin price prediction: A case study yahoo finance stock market. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019. IEEE, 206-210.

GURRIB, I. & KAMALOV, F. 2021. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies in Economics Finance*.

HANNA, A. J., TURNER, J. D. & WALKER, C. B. 2020. News media and investor sentiment during bull and bear markets. *The European Journal of Finance,* 26**,** 1377-1395.

HOTZ-BEHOFSITS, C., HUBER, F. & ZÖRNER, T. O. 2018. Predicting crypto-currencies using sparse non-Gaussian state space models. *Journal of Forecasting,* 37**,** 627-640.

IQBAL, M., IQBAL, M. S., JASKANI, F. H., IQBAL, K. & HASSAN, A. 2021. Time-series prediction of cryptocurrency market using machine learning techniques. *EAI Endorsed Transactions on Creative Technologies***,** e4.

JAIN, A., TRIPATHI, S., DWIVEDI, H. D. & SAXENA, P. Forecasting price of cryptocurrencies using tweets sentiment analysis. 2018 eleventh international conference on contemporary computing (IC3), 2018. IEEE, 1-7.

JANA, R., GHOSH, I. & DAS, D. 2021. A differential evolution-based regression framework for forecasting Bitcoin price. *Annuals of Operations Research,* 306**,** 295-320.

JIANG, Z. & LIANG, J. Cryptocurrency portfolio management with deep reinforcement learning. 2017 Intelligent Systems Conference (IntelliSys), 2017. IEEE, 905-913.

KAGGLE. 2022. *Kaggle* [Online]. Available: https://www.kaggle.com/ [Accessed 30th May 2022].

KARALEVICIUS, V., DEGRANDE, N. & DE WEERDT, J. 2018. Using sentiment analysis to predict interday Bitcoin price movements. *The Journal of Risk Finance*.

KHURANA, D., KOLI, A., KHATTER, K. & SINGH, S. 2017. Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:.05148*.

KILIMCI, Z. H. 2020. Sentiment analysis based direction prediction in bitcoin using deep learning algorithms and word embedding models. *International Journal of Intelligent Systems Applications in Engineering,* 8**,** 60-65.

KRAAIJEVELD, O. & DE SMEDT, J. 2020. The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions Money,* 65**,** 101188.

LI, Y. & DAI, W. 2020. Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model. *The journal of engineering,* 2020**,** 344-347.

LINARDATOS, P. & KOTSIANTIS, S. 2020. Bitcoin Price Prediction Combining Data and Text Mining. *Advances in Integrations of Intelligent Methods.* Springer.

LY, B., TIMAUL, D., LUKANAN, A., LAU, J. & STEINMETZ, E. Applying deep learning to better predict cryptocurrency trends. Midwest Instruction and Computing Symposium, 2018.

MITTAL, A., DHIMAN, V., SINGH, A. & PRAKASH, C. Short-term bitcoin price fluctuation prediction using social media and web search data. 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019. IEEE, 1-6.

MOHAPATRA, S., AHMED, N. & ALENCAR, P. Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments. 2019 IEEE International Conference on Big Data (Big Data), 2019. IEEE, 5544-5551.

MUDASSIR, M., BENNBAIA, S., UNAL, D. & HAMMOUDEH, M. 2020. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. *Neural Computing and Applications***,** 1-15.

NAGPAL, A. & GABRANI, G. Python for data analytics, scientific and technical applications. 2019 Amity international conference on artificial intelligence (AICAI), 2019. IEEE, 140-145.

NAKAMOTO, S. 2008. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review***,** 21260.

PAGOLU, V. S., REDDY, K. N., PANDA, G. & MAJHI, B. Sentiment analysis of Twitter data for predicting stock market movements. 2016 international conference on signal processing, communication, power and embedded system (SCOPES), 2016. IEEE, 1345-1350.

PANO, T. & KASHEF, R. 2020. A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. *Big Data Cognitive Computing,* 4**,** 33.

PANT, D. R., NEUPANE, P., POUDEL, A., POKHREL, A. K. & LAMA, B. K. Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), 2018. IEEE, 128-132.

PASCALL, V. K. 2020. *Contractions* [Online]. Github. Available: https://github.com/kootenpv/contractions [Accessed 2nd November 2021].

PIENAAR, S. W. & MALEKIAN, R. Human activity recognition using LSTM-RNN deep neural network architecture. 2019 IEEE 2nd wireless africa conference (WAC), 2019. IEEE, 1-5.

POONGODI, M., SHARMA, A., VIJAYAKUMAR, V., BHARDWAJ, V., SHARMA, A. P., IQBAL, R. & KUMAR, R. 2020. Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. *Computers Electrical Engineering,* 81**,** 106527.

PRADHA, S., HALGAMUGE, M. N. & VINH, N. T. Q. Effective text data preprocessing technique for sentiment analysis in social media data. 2019 11th international conference on knowledge and systems engineering (KSE), 2019. IEEE, 1-8.

PRAKASH, T. N. & ALOYSIUS, A. 2019. A Comparative study of Lexicon based and Machine learning based classifications in Sentiment analysis.

RATHAN, K., SAI, S. V. & MANIKANTA, T. S. Crypto-currency price prediction using decision tree and regression techniques. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019. IEEE, 190-194.

RIQUELME, F. & GONZÁLEZ-CANTERGIANI, P. 2016. Measuring user influence on Twitter: A survey. *Information processing management,* 52**,** 949-975.

SIAMI-NAMINI, S., TAVAKOLI, N. & NAMIN, A. S. The performance of LSTM and BiLSTM in forecasting time series. 2019 IEEE International Conference on Big Data (Big Data), 2019. IEEE, 3285-3292.

SIBEL KERVANCI, I. & FATIH, A. 2020. Review on Bitcoin Price Prediction Using Machine Learning and Statistical Methods. *Sakarya University Journal of Computer Information Sciences,* 3**,** 272-282.

STELZMÜLLER, C., TANZER, S. & SCHEDL, M. Cross-city Analysis of Location-based Sentiment in User-generated Text. Companion Proceedings of the Web Conference 2021, 2021. 339-346.

STENQVIST, E. & LÖNNÖ, J. 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

ULUMUDDIN, I., SUNARDI, S. & FADLIL, A. 2020. Bitcoin Price Prediction Using Long Short Term Memory (LSTM): Bitcoin Price Prediction Using Long Short Term Memory (LSTM). *Jurnal Mantik,* 4**,** 1090-1095.

URAS, N., MARCHESI, L., MARCHESI, M. & TONELLI, R. 2020. Forecasting Bitcoin closing price series using linear regression and neural networks models. *PeerJ Computer Science,* 6**,** e279.

VALENCIA, F., GÓMEZ-ESPINOSA, A. & VALDÉS-AGUIRRE, B. 2019. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy,* 21**,** 589.

WOŁK, K. 2020. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems,* 37**,** e12493.

WOOLEY, S., EDMONDS, A., BAGAVATHI, A. & KRISHNAN, S. Extracting cryptocurrency price movements from the reddit network sentiment. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019. IEEE, 500-505.

YUE, L., CHEN, W., LI, X., ZUO, W. & YIN, M. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems,* 60**,** 617-663.

Appendices

Appendix A

Environment:

Software:

Package	**Use Case**
Pandas	It is mostly used for data handling and data manipulation. Pandas library was used for handling the data and also being able to read and write CSV files.
NumPy	This library was used for mathematical operations on arrays.
Matplotlib	Visualization of plots. In this study, it was used for displaying line charts and doughnut charts
NLTK	NLTK library is specifically for NLP tasks and it was used for providing the Twitter samples, TextBlob and pre-processing steps.
Langdetect	To detect non-English tweets.
vaderSentiment	The VADER sentiment analyser was used to extract the polarity from users' opinions.
Tensorflow	Setting the seed for training and testing
Keras	To build neural network models such as LSTM and BILSTM.
Seaborn	Visualization of correlation heatmaps.
DateTime	Utilised to handle and manipulate date and time data types.
sklearn	This library provides efficient tools for machine learning and was used for data normalization, metrics, splitting the data and confusion matrices.
Re	This library stands for Regular Expression operators known as Regex which were used for preprocessing steps.

Table A.1: Depict the Python packages utilised throughout this research.

Hardware:

Machine	CPU	RAM
HP-Laptop 15-bs591nd	Intel Core i5-7200U CPU @ 2.50GHz-2.70 GHz	16GB

Table A.2: Hardware used throughout this research.

Appendix B

To view the source code of this project a GitHub repository link is provided:

https://github.com/RyanPirotta/RyanPirotta_Dissertation.git