C O V E N T R Y

U N I V E R S I T Y

Faculty of Engineering, Environment and Computing

School of Computing, Electronics and Mathematics

MSc. Data science and computational Intelligence

7151CEM – Computing Individual Research Project

# LEVERAGING TWITTER SENTIMENT ANALYSIS TO ANTICIPATE BITCOIN VALUE

Author: Vamsidhar Reddy Muchurami

SID: 10864627

SUPERVISOR: SIMON BILLINGS

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Computational Intelligence

Academic Year: 2021/2022

# Declaration of Originality

I declare that this project is all my work and has not been copied in part or whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, the internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for plagiarism prevention and detection.

# Statement of copyright



# Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed:
Vamsidhar Reddy Muchurami
Date:
23-07-2022

Please complete all the fields.

| | |
|---|---|
| First Name | Vamsidhar Reddy |
| Last Name | Muchurami |
| SID | 10864627 |
| Ethics Application Number | P137852 |
| 1st Supervisor Name | Simon Billings |
| 2nd Supervisor Name | Furrkh Aslam |

**This form must be completed, scanned, and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.**

# Abstract

Twitter sentiment has demonstrated that it can be used to forecast whether the price of bitcoin will rise or fall. However, we can only forecast the price direction, not the degree of growth or drop. To estimate the direction and amount of the growth or drop, we want to expand the research in this study. We use the number of tweets and sentiment analysis by users. To determine the ideal time at which the sentiment stated becomes a trustworthy signal of price changes, we provide findings from studies examining the relationship between future price differentiation.



# Acknowledgements

I would like to express my heartfelt appreciation to Prof. Simon Billings, my supervisor, for his direction and assistance in establishing this great research and project. I would want to express my gratitude to Coventry University for the opportunity to share my thoughts and research with the world.

I am indebted to my parents for allowing me to study at Coventry's excellent University. I've learnt a great lot here that has aided me in living a successful life, which would not have been possible without my parents' support. I should also express my gratitude to my brother, Sushmita my friends for their assistance throughout the project's most trying phases. Many thanks to everyone who assisted in bringing out the best in me.

# 1. Introduction

The word Bitcoin was described in a white paper published in 2008. Bitcoin (₿) (Wikipedia contributors. et al., 2022, July 3) is the decentralised digital currency which is peer-to-peer digital money that can be transferred over the Defi network. Later, the transactions are cryptographically verified by network nodes and stored in public databases as distributed ledgers known as the blockchain. The cryptocurrency was first created in 2008 by an unknown individual or group of individuals under the name of Satoshi Nakamoto (4 Key Concepts of Bitcoin - GeeksforGeeks. et al., 2021).

Machine learning is another subset of Artificial Intelligence(AI), which allows for interpretation and provides meaning to big data to address real-world problems (Moné, L. et al., 2022). For example, in Bitcoin price prediction (Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. et al., 2020) using previous data in the analysis of various impacts of price volatility and rise or fall or constant movement in value or trend. In another aspect machine learning is part of teaching the computer to learn using statistical techniques or computer models.

There is an ongoing discussion in public and academics regarding what determines Bitcoin's worth and if it is anything more than hype or a bubble (Using sentiment analysis to predict interday Bitcoin price movements | Emerald Insight. et al., 2022). In this context, there are several research possibilities to better understand the determinants of the Bitcoin price. Previous research concentrated on fundamental and technical analysis to forecast price fluctuations and uncover casual links. Furthermore, there are few scholars and reports that investigate the broader social elements of Bitcoin price using sentiment. This research will continue attempting to ascertain if the value of bitcoin is mostly determined by sentiment rather than more fundamental drivers or factors.

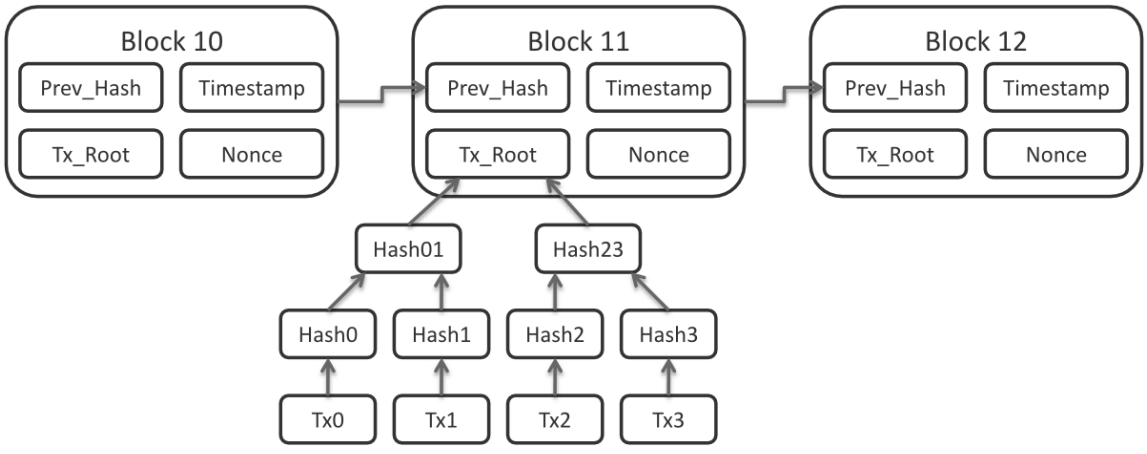


Fig 1: Structure of Bitcoin (₿)

## Background To The Project:

Bitcoin was created as a project to support internet money, the digital currency for the world using hash rates which will be verified by other users or miners. Bitcoin is stored in different servers all around the world and is based on the SHA-256 algorithm which is designed by the US national security agency.

It is frequently considered what type of asset class cryptocurrencies belongs to. Although they are debated as the sense of a digital medium of exchange they have restrictions. Individuals care to use official currencies such as US Dollar (USD) or Euro (EUR) or British Pound Sterling (GBP) because they are stable and are monetized by governments. Cryptocurrencies do lack these features as they are maintained by no authority and are highly volatile causing the market to render crypto as an unsuitable medium to store or exchange assets (The economics of Bitcoin price formation et al., 2022).

Few countries made bitcoin a legal tender whereas Ukraine is taking bitcoin donations to aid for Russian invasion (Bloomberg et al.,2022). However, it has been more than 1.5 decades since the bitcoin was invented but there is still an issue even in the advanced countries like the USA, UK, etc., where people still feel is it safe for them to invest in bitcoin or is it just a hype for the sharks and big investors to make money by promoting it.

An article in (Bloomberg - Are you a robot?) (Matt Levine et al., May. 17, 2022) contradicts that Elon musk using his social media account is tweaking the bitcoin and other cryptocurrencies at his will. Another article in Forbes (Bambrough, B. et al.,2022) supports the statement above as he was hit by a huge lawsuit for using social media platforms to change people's sentiment towards bitcoin and dogecoin. As Elon tweets a post about bitcoin positively saying Tesla (TSLA) will accept bitcoin as legal tender the price goes up and when he posts a negatively saying bitcoin is so much harm to the planet as it is using powerful computers which run on electricity and are draining the natural resources the price will go down.

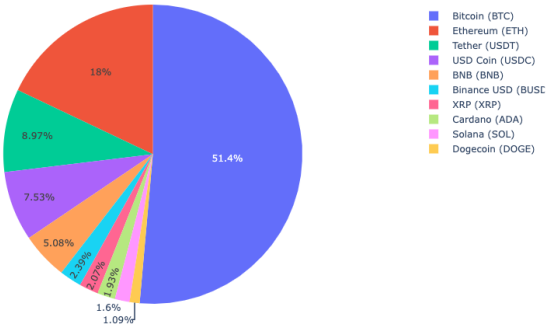


Fig-2: Pie chart percentage of Top 10 cryptocurrencies

This might be because the power of social media had a raffling towards the crypto market now dogecoin has lost 90% of its market cap, and bitcoin lost from 1.2 Trillion to 398 Billion USD. Some trusted bitcoin is the new gold against inflation, As the inflation raises year by year bitcoin is plummeting up and down and it is important to check how people's opinion is with the bitcoin.

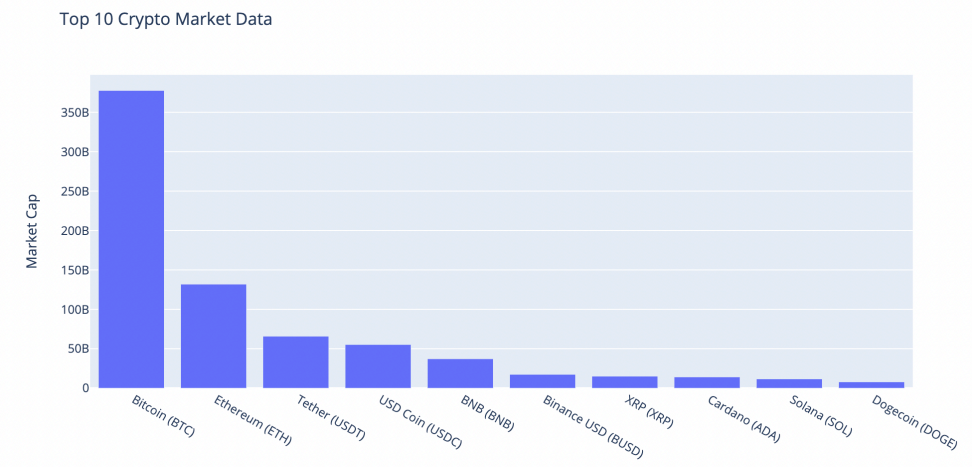


Fig-3 Top 10 crypto against billion dollars

## Project Objectives:

As technology advances in the 21st century, the currency is changing to decentralized networks. Bitcoin is the cryptocurrency that came into existence to effectively challenge and help people with transactions around the world without any hassle. Bitcoin is the type of currency that helps people to store and transact their wealth without any 3rd party. Though many investors and businesses are backing crypto assets it is considered to be highly volatile despite their contribution to society.

The objective of this paper is to identify whether Twitter sentiment is affecting the bitcoin price by using different principles of data science using machine learning models for prediction. The study may help policymakers or bitcoin investors to reduce the risk of huge losses of money.

### Intended User or group and their requirements:

In this thesis, by recognising the issue using social media like Twitter how people's opinions on Bitcoin price are increasing by the force of only a single entity and/or the huge number of people boosting the price. This will help key investors' behaviour in investing or selling the stock(Raviv, P M et al., 2018). The vision of this project is to use statistical analysis and data mining techniques to conclude social media influence and bitcoin price.

### Model requirements:

Any project will start with a set of processes for a client or user. It is as follows,

a. Collection of data, it might be downloading the data from using different types of APIs or finding a good dataset in websites or repositories.
b. Analyse the data and finalise it.
c. IDE installation.
d. Install packages for the project.

## Overview of the Report:

This section is to give overview of report and methodologies I followed,

The report started with declaration, statement, originality of the project followed by abstract which explains the overview of the whole project with aim and conclusions.

First section is introduction to the problem statement and evolution of bitcoin, bitcoin flaws and advantages. The future of the bitcoin might depend on twitter sentiment with some of the methodologies. A time-series of both datasets is considered to check for correlation. As detailed to the problem strategies solution to be found out in later sections.

Second section is the literature review of the project, which mainly refers to the previous papers, methodologies, and ideas followed. This is a comprehensive idea of solutions to the sentiment finding in the time-series data which previous papers not able to find. By doing this we will have a good idea of which type of analysis to do and what to avoid. This will allow us to find the good results and cover the topics.

Next section is methodology in this study, this describes the challenges, techniques, finding the dataset, pre-processing the data, and cleaning the dataset to get it ready to the analysis, training, and testing of dataset for modelling, find the perfect model like which model and method to use.

Next section is requirements, which describes about the non-functional and functional requirements for the project.

Next section is data analysis, which describes the analysis methods and plotting or visualizations of the data to give good idea of the dataset.

Next section is implementation, which describes techniques, modelling of the dataset.

Next section is training, describes the training of a dataset to build a model and compare them for results.

Next section is testing, describes the testing the model to find the relation of the data.

Next section is project management, describes the project schedule, quality management, risk assessment, etc., which are followed to achieve this project.

Next section is critical appraisal

Next section is conclusion, which describes the results of the model and future work that might be done.

Next section is student reflections, the experiences and challenges faced during and before the project.

The final section is bibliography and references, which includes references and studies I followed to make this project.

# 2. Literature Review

This section involves the past reviews and background or related works of the project, the cryptocurrencies are discussed on the basis followed by Twitter sentiment analysis and its role in the financial markets.

The authors of “Sentiment Analysis of Twitter Data for Predicting Stock Market Movements” (V. Pagolu, K. Challa, G. Panda et al., 2016) employed social media microblogging to estimate stock prices since it presents the general public opinions and views about current events. They explored the relationship between a company’s stock price and tweet sentiment using sentiment analysis by supervised machine learning techniques. The report shows there is a definite relation between the increasing and decreasing value of a stock.

The authors of “Prediction of Bitcoin Price using Deep Learning Model” utilized yahoo finance data of 7 years to get this value, they used deep learning models like Recurrent Neural Networks(RNN) (S. E. Freeda, T. C. E. Selvan and I. G. Hemanandhini et al., 2021). They compared with various machine learning models such as Gaussian naïve Bayes, SVM, random forest, and K-Nearest neighbours. They had got 76.99% when compared to other models. While the accuracy got better but the log loss was 7.18%.

A study on “Bitcoin price prediction” focused on predicting the Bitcoin on certain criteria or attributes that impact the bitcoin price (L. Felizardo, R. Oliveira, E. Del-Moral-Hernandez, and F. Cozman, et al., 2019). The researchers conducted an extensive study on other factors that affect the accuracy of the price prediction and now they are selecting factors that impact the price. They used support vector machines to analyse the data in an 80:20 ratio. The main advantage of SVM is that they have a high space complexity and a low time complexity.

In “Bitcoin Spread Prediction Using Social and Web Search Media” (M. Matta, I. Lunesu, M. Marchesi et al., 2015) explored a dataset of 1.9 million tweets to check the value of bitcoin. They compared the price trends to google trends data, and tweet volume and checked tweets polarity if the tweets are positive or negative. They evaluated the tweets by giving them scores for each tweet and labelled them. They discovered the correlation between bitcoin’s price, positive response and google trends. Related to this “Predicting Bitcoin price fluctuation with Twitter sentiment analysis” (E. Stenqvist, J. Lönnö, et al., 2017) studied whether user’s tweets are making an impact on price fluxes by analysing more than 2 million tweets related to bitcoin within one month, they reported that applying machine learning to the study there might be a connection.

In “Using time-series and sentiment analysis to detect the determinants of bitcoin prices” (Kraaijeveld, O., & De Smedt, J. et al., 2020) Twitter sentiment analysis by georgoula used SVM regression models to forecast the fluctuations in their study (Georgoula et al., 2015). They achieved 89.6 percent of accuracy and discovered just a short-term association between Twitter sentiment is positive on bitcoin. Another study (Garcia and Schweitzer, et al., 2015) found that Twitter sentiment polarity follows price changes using a lexicon-based method with granger-causality testing and a vector autoregressive(VAR) model. Another study (Mai et al. 2015) combined intraday analysis and demonstrated that tweets may be used to anticipate bitcoin prices.

In “The use of artificial neural networks in the analysis and prediction of stock prices” (F. A. d. Oliveira, L. E. Zarate, M. d. A. Reis and C. N. Nobre, et al., 2011) while anticipating the stock’s price, the trends had a pattern which performed in short-term (weeks) and long-term (Months) and found out there was a 79% of accuracy with tweets. The performance evaluation metrics evaluated to buy or sell the stock based on the projected output using sentimental analysis.

Similar methodologies have been used in research that forecasted, where a sentiment analysis methodology is frequently paired with a (Granger_) causality test and/or a regression model to clean their Twitter data, researchers used sophisticated pre-processing techniques such as stop-word removal, stemming, tokenisation, and filtering out non-English tweets. Many of the publications that adopt a lexicon-based approach make use of the Loughran and McDonald financial corpus(Mai et al. 2015, Li et al., 2014, Karalevicius et al., 2018).

Another analysis “Bayesian regression and Bitcoin,” (D. Shah and K. Zhang, et al., 2015) discussed time series data of 10 minutes and 10 seconds time intervals. They prepared 3 time-series datasets for 30, 60, and 120 minutes. Then performed a random forest or GLM approach on the dataframe and created 3 linear models to forecast the price of bitcoin.

However, this study was constrained by the fact that the bitcoin prices were obtained in the form of only one exchange or social media like Twitter is making a part to go prices go up and down.

# 3. Methodology:

This section mainly focuses on gaining deeper insight into challenges, techniques, and the type of methodology followed to develop the project. Agile methodology has been chosen for this project since it is an iterative method to follow and relies on repeated tasks to get good results. This differs from general approaches where you need to follow a corporate approach where you need to develop according to designated requirements and can’t change them in the iterative testing and rapid prototype environment. The agile method is considered a dynamic approach favourable to growth across all the available resources. The goal of agile is to create an environment to create a plan that aligns with organization goals while also offering a framework for continuous integration, development, testing and development.


When adopting agile methodology, models and data are prioritised based on the project’s goals and needs. This will result in the possible outcome and time for troubleshooting and changes in the code to meet requirements.

Because the data is constantly changing, it is difficult to anticipate the time and effort required to execute a data science project. There will be room for continuous improvement in the development process. This is why agile methodology is suitable for most data science projects. This methodology will help organise the project structure and management if we employ methods like Scrum or Kanban methods.

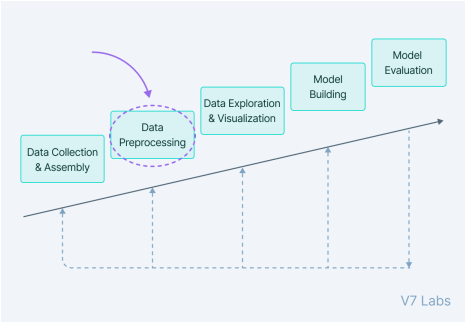


Fig-4 Steps in ML model building

## Dataset:

Choosing the dataset is a big task in the project, which is a mix of different attributes. Data will be arranged in various formats .csv, .xlsx, .json etc., it is organised in so many ways with challenges. The dataset I chose is from Kaggle which is a repository for many datasets to practice on for doing analysis. Twitter sentiment on the bitcoin dataset has tabular columns and rows with attributes. A time-series dataset from 2021-02-05 to 2022-01-22 is used for this project to show visualisations in this study. Time series data is univariate and simple to utilise for analysis to show the historical behaviour of sentiment for forecasting. Dataset has 2.2 million rows, and 13 columns as follows:

user_name – username of the Twitter user
user_location – location from where the user is tweeted
user_description – description of a tweet
user_created – when was the user registered on Twitter
user_followers – how many followers does the user have
user_friends – how many users does he follows on Twitter
user_favourites – how many favourited the user
user_verified – user is verified or not
date – tweet's date
text – actual tweet from Twitter
hashtags – hashtags used in the tweet
source – from which source the user is posting it
is_retweet – is it re-posted again

## Data Pre-process

Pre-processing of a dataset is the crucial step in any data science project because however the data is collected either from the APIs or from repositories it will not be of good quality or will have too much-undesired data. The error includes defects in the dataset, human errors, data gathering techniques, NaN values, and the problem with the collection process or code which will affect the creation of a good model. To create a good ML model, it is important to clean the dataset in pre-process. This process "Data Pre-processing in Machine Learning [Steps & Techniques] et al., 2022)" includes identifying NaN values in the dataset, converting data types, finding duplicates to drop them or update missing values, normalizing data and outliers which will disturb model performance which may lead to giving false predictions. This is a repetitive process as shown below, we will explore following the report later.


Fig-4 Cycle of Data Processing

- Data pre-process includes data cleaning,
- Missing values – identify to either update them or remove them.
- Noisy data – clustering, regression, binning.
- Outliers – normalize data.

## Train & Test Data

After pre-processing, the dataset will be used to train a model using python modules like Sklearn, TensorFlow etc., to create train and test data split. This section briefly discusses splitting and the significance of data splitting. Simply to explain this, to develop an ML model a dataset with huge data is sent and split into two ratios then the data is fed into train the model to predict. To train and test the model, a dataset is divided into two parts train, test with ratios of 70:30 or 80:20 or any desirable split. The training dataset will be larger than the test dataset because they are trained to the model and test data is to evaluate the model accuracy.

## Machine Learning models

ML models are used to teach the computer of statistical approaches, these are classified into several categories based on the type of model we are building. There are many modules which we can use to train datasets or pre-trained models like regression models, classification models, and clustering models etc., In our case, we are going to build a model to predict the sentiment of people. We have to use sentiment-based modules such as Vader sentiment and LSTM. These two models are used for text-based modelling or natural language processing (NLP).

The natural language process (NLP) is a subfield of the artificial intelligence branch which involves obtaining insights from human language and interpreting text and voice. Such applications are language translation, voice recognition, virtual assistants, chatbots, and other typical applications. Real-world applications are,

- autocompletion in phones and computers.
- Siri, Alexa, and google assistant.
- Spam detection.
- Document categorisation.

We need to understand the NLP because in today's world there are 2.5 quintillion bytes of data being generated every day and the bulk of them is unstructured like text, audio, and video on different websites. NLP is one of the techniques to extract and understand that will give meaning to the huge data. There are five important stages in NLP,

### Lexical Analysis

It is the first step of an NLP project, in which we break the text and tokenize the words. It will remove unnecessary blank/white spaces.

### Syntactic Analysis

This step involves examining how words are ordered in a sentence to determine if they are in the correct order. It also includes determining if the sentence is grammatically accurate or not. Stemming or Lemmatization is the most common technique used in this process. Stemming does not always have to result in a meaningful term, but lemmatization would give the meaningful term.

**Semantic Analysis**

This step involves searching the dictionary for the meaning and checking if the words are meaningful.

**Discourage Integration**

This step involves determining the meaning of the sentence in any paragraph. In this section, we look at how the immediate occurrence of words/sentences affects the subsequent words followed.

**Pragmatic Analytics**

This is the final stage of the NLP process, which involves extracting insights from the text input and understanding the context.

**LSTM**

We need to understand the RNN model to better understand LSTM, RNN is a supervised deep learning model. The neurons are linked to one over another time in this case “NLP | Sentiment Analysis using LSTM - Analytics Vidhya. et al., 2021)” the idea behind the RNN is to remember the first information that is generated and pass it to the further analysis in the future. It signifies that data from one instance t1 is used as input for the following instance t2.

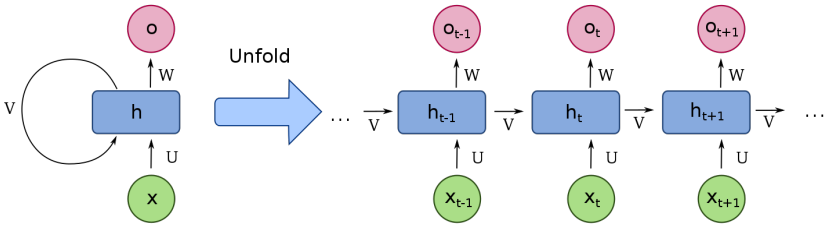


Fig-5 LSTM architecture

New weight = Old weight – (learning rate * gradient)

One of the issues is the RNN is a vanishing gradient. During the training phase of any neural network, the weights are changed by computing the error and back-propagation. However, it is rather hard in the case of RNN “NLP | Sentiment Analysis using LSTM - Analytics Vidhya. et al., 2021” since we need to propagate overtime to the neurons. The problem lies in calculating the weights. The gradient calculated at each time is then multiplied by the weights computed previously. As we go back in through time in the network for calculating weights the gradient descends, causing the gradient to disappear. If the gradient value is really small then it will not impact the learning process.

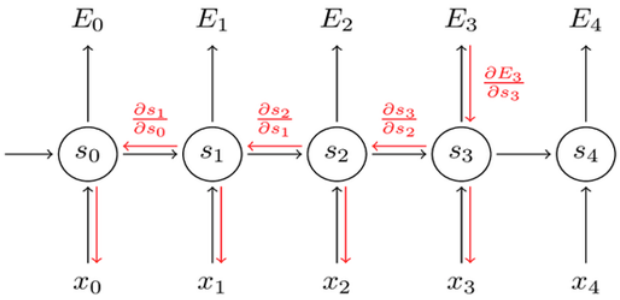


Fig-6 backpropagation Through Time

From Fig-6, assume that we estimated the error at the time t3 instance is calculated. To update the weights of all neurons that have t3 time instances, we need to propagate through instances t0. If the sequence is long, the vanishing gradient will make it impossible for RNN to transmit information from one occurrence to another. LSTM model is used to overcome this problem, it is an updated model to RNN to address the vanishing gradient problem. Fig-7 is the architecture of how information is carried inside the model.

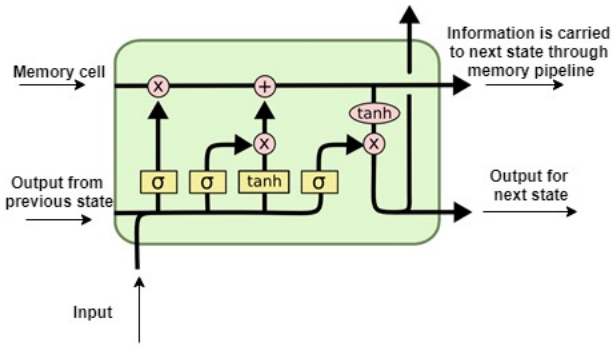


Fig-7 Architecture of LSTM

It features that has a memory cell to carry information from one-time instance to the next. When compared with RNN it can recall from previous and feed it to the next and avoid vanishing gradient problem. This will help to add or remove the memory cell with the help of valves. LSTM network is fed on the current data instance from the data and hidden layers output from the first instances. The two data outputs will pass through different activation values and functions in the network.

**Vader sentiment**

Sentiment analysis is the process of detecting emotions using positive, negative, or neutral in a text using machine learning “Python | Sentiment Analysis using VADER - GeeksforGeeks. (2019)”. It is also known as opinion mining where it involves determining the viewer's point of view (POV) or feelings towards a topic, technology, or business model etc., the question here is why we perform sentiment analysis?

Businesses use sentiment analysis to establish strategies, analyse customers' opinions on brands or products, or how people are reacting to campaigns or new launches and why they buy them. In Politics they keep track of political agenda management by analysing public viewpoints and to find consistency and discrepancy between actions and government policy statements, it may be used to forecast the results of elections. Sentiment analysis is also used

to track and analyse social phenomena, identify potentially harmful events, and evaluate the general mood.

**VADER Sentiment Analysis**

**VADER** (Valence aware dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that maps lexical information to emotion intensities known as scores "SENTIMENTAL ANALYSIS USING VADER et al., 2020". It is sensitive to both polarity (positive/negative) and the intensity (strength) of emotion. It is included in the NLTK package of python and may be used on unlabelled text data directly. A text's score is calculated with sentiment to the intensity of each word in a tweet. For example, words like love, to the moon, great etc., also Vader understands negative sentiment words like bad, worse, and fall etc., it is important to recognise the importance of capitalization and punctuation such as "GOOD".

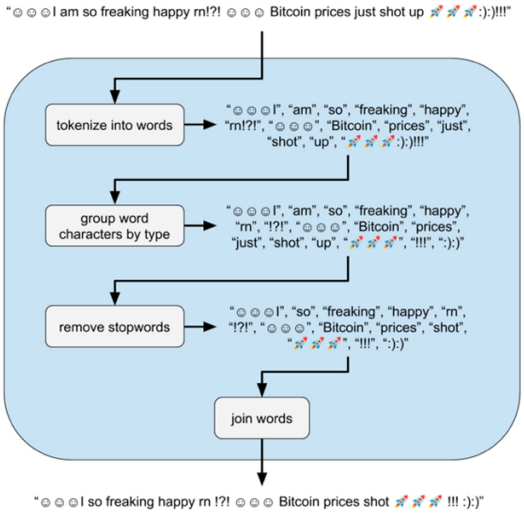


Fig-8 Vader architecture

**Polarity Classification**

We will not try to identify if a text is subjective or objective, fact, or opinion rather we are solely interested in whether the language is represented in positive, negative, or neutral. We will try to combine all sentences in a document or paragraph to form an overall view.

**Document Level Scope**

We will also attempt to aggregate all of the sentences in a document or paragraph to form an overview.

**Broad Steps**

First, consider a text to be analysed. A model trained on paragraph-length reviews might not be useful. We need to use a model that is suitable for the job at hand. Next, choose which type of analysis we have to perform for good results. Some basic sentiment analysis will go further and be handled with two-word combinations known as bigrams. We will start with entire sentences, and we will use lexicon-based Vader to add scores for the text.

# 4. Requirements:

Requirements are a set of tasks or activities that must be met to execute a project.

### Functional Requirements

Functional requirements are functions or features that developers must follow to complete their user's product requirements. It is critical to meet both user and developers' functional requirements to maintain data, show visualisations, and understand data.

Below are the specifications of the computer I used to develop the model. Hardware:

- Model Name: Apple MacBook Air
- Model Identifier: MacBook Air 10,1
- Chip: Apple M1
- Total Number of Cores: 8 (4 performance and 4 efficiencies)
- Memory: 8 GB

Software used:

- Python, jupyter-notebook, Vader, Sklearn, matplotlib

**Non-Functional Requirements**

- Repository
- Labelling every section
- Write .readme file and documentation
- Code standard

# 5. Analysis:

The analysis is a part of a data science project where we try to understand the data better by either visualisation, printing data etc., the process will come down to a small bit to gain good knowledge. A comprehensive investigation of data to grasp and discover its features.

In our case, a detailed observation of the bitcoin price close should be understood by visualisations. Fig-9 shows the close value of the bitcoin price against time, this graph shows the bitcoin price has been increasing exponentially from 2018 to 2021 and there is a fall again in the mid of 2021.

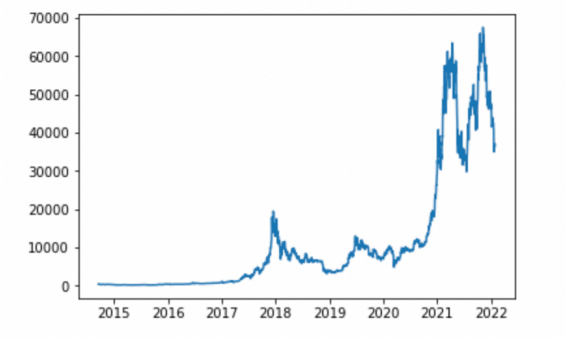


Fig-9 Bitcoin value overtime closing time

We can also observe that it's been slowly improving from 2017 to 2020 whereas there is an exponential growth in the years 2019 to 2022.

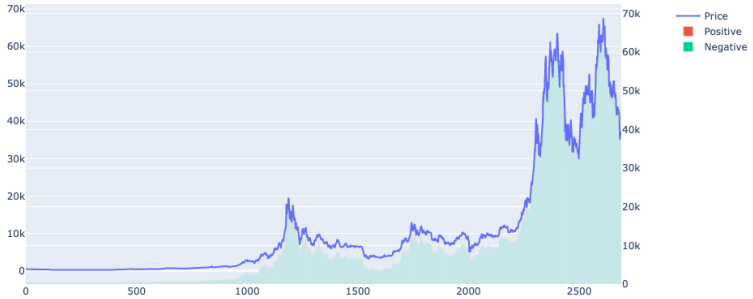


Fig-10 Growth Of Bitcoin Price

Here in the fig-10 we can see the market close specifically growing from 0 to 69000 USD from the start of bitcoin. There is a good growth in value from past years always closing in positive.

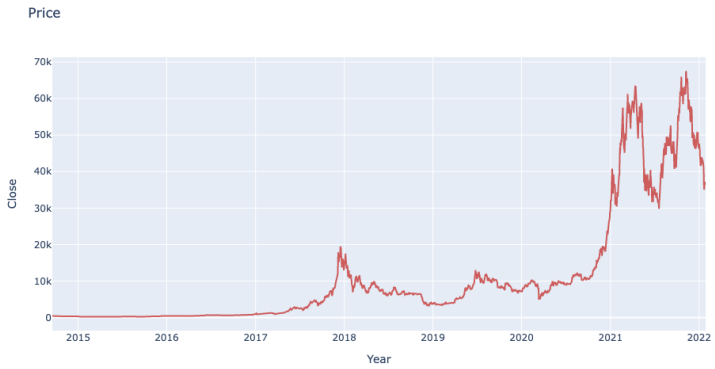


Fig-11 Growth of Bitcoin Years VS Close

In the fig-11 we can observe that bitcoin started performing well in specific periods years 2018, 2020, and 2021 and there is a subsequent fall in the middle years of 2021 to 2022. It clearly shows there is a market cap fall in the years 2021 and 2018.


| | Date | Open | High | Low | Close | Adj Close | Volume | tgl |
|---|---|---|---|---|---|---|---|---|
| **2609** | 2021-11-08 | 63344.066406 | 67673.742188 | 63344.066406 | 67566.828125 | 67566.828125 | 41125608330 | 67566 (08 November) |
| **2610** | 2021-11-09 | 67549.734375 | 68530.335938 | 66382.062500 | 66971.828125 | 66971.828125 | 42357991721 | 66971 (09 November) |
| **2590** | 2021-10-20 | 64284.585938 | 66930.390625 | 63610.675781 | 65992.835938 | 65992.835938 | 40788955582 | 65992 (20 October) |

Fig-12 Bitcoin Close at High

In the fig-12 we can see the highest close of bitcoin is on 3 dates in the whole period of 2015 to 2022 and the close is at 67.5k on 2021-11-08 USD which is the highest in bitcoin history but there is the bear market which followed panic in investors which resulted in mass selling of bitcoin in investors substantially making it lower to 20k USD now. It is important to see how people are reacting to the price of bitcoin sentiment through Twitter or are the people boosting prices by reacting positively at this time.

# 6. Implementation:

## Model Building

In this section, we will look in deep at the process of creating an ML model using packages from python. There are a few steps in the process to achieve good results as follows “How To Develop a Machine Learning Model From Scratch et al ., (2021)”.

- Problem Statement(objective, desired outputs…).
- Gathering data.
- Measuring success.
- Set an evaluation protocol and the different protocols available.

- Data Preparation
- Split the data.
- Differentiate between over and underfitting.
- An overview of model learning outcomes.
- Regularization.
- Develop a model.
- Choose an adequate model and tune it to the best performance possible.

## Language and Platform

Python is the programming language used in this project and I used jupyter-notebook as IDE. There are built-in packages used from python to perform EDA, reading datasets, model building and Machine learning.

### a. Python for ML

In this project, we are using an open-source language such as python. The methodology for the project is open-source and easily understandable as python is a language which follows the OOPS concept. We can easily use this because It is written in an object-oriented approach, code readability, and embedding etc., there are many libraries which will help in building ML models, performing analysis, and creating visualisations such as Keras, TensorFlow, matplotlib, pandas and so on. There are several libraries available for creating visualisations as well.

### b. Jupyter-notebook

I used jupyter-notebook as IDE for the project because it is a browser-based interactive development environment that can execute cell by cell. This will help in debugging the code and understanding the data to perform analysis and create interactive graphs it is free to use, and we can also share the code with .ipynb extensions.

There are other IDEs such as Spyder for python, atom, sublime-text, and the most famous Microsoft VS code.

## EDA

Exploratory data analysis is the process of doing preliminary analysis on data to understand it better to get patterns, find outliers, normalize the data, test hypotheses, and validate using statistics and graphical representations. Some of these analyses give a clear idea of data and make models and some will have a dead end. EDA is an important part of a data science project, even if we have a dataset from repositories like Kaggle because we need to check the quality of the dataset before performing our core analysis.

One of the processes in EDA is data cleaning, to check whether data meets requirements to create visualisations, transformation, and modelling.

In fig-13 we can see that bitcoin is effectively increasing from 2021 march to 2022 January and the green label indicates that people interacting on Twitter either positive or negative to

boost or fall its performance. There is a good growth in the period from 2021 July to 2021 November which resulted in a higher close than the previous close. But there might be outliers in this data, and first, we need to perform EDA on the dataset to normalize it.

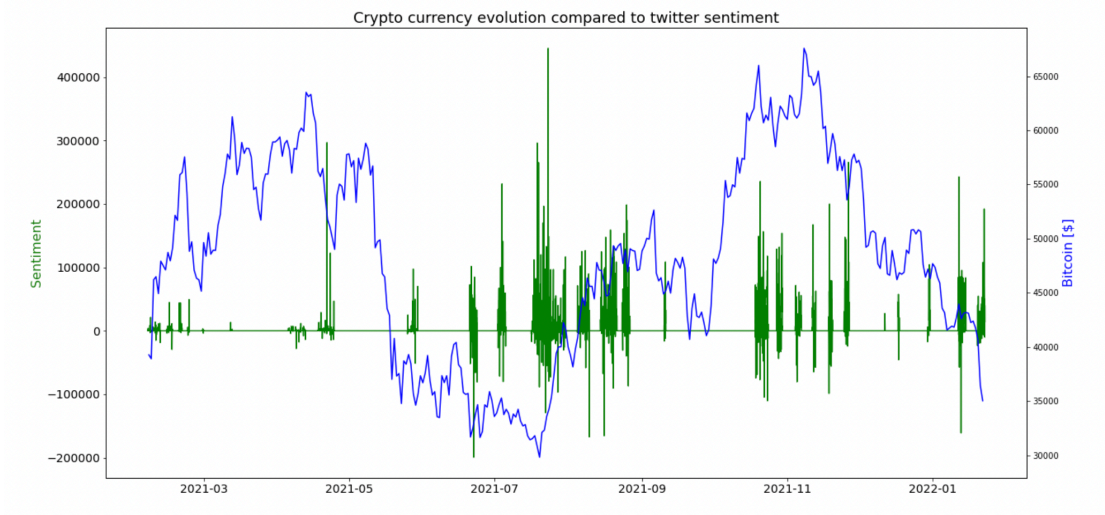


Fig-13 Bitcoin Price Change With Sentiment without normalized data

**Data Pre-Process**

The most important feature of the ML model is collecting data and pre-processing it. There are several methods to collect data from a source using APIs, finding from repositories or web scraping.

| user_location | user_description | user_created | user_followers | user_friends | user_favourites | user_verified | date | text | hashtags | source | is_retweet |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Atlanta, GA | Biz Consultant, real estate, fintech, startups... | 2009-04-26 20:05:09 | 8534.0 | 7605 | 4838 | False | 2021-02-10 23:59:04 | Blue Ridge Bank shares halted by NYSE after #b... | ['bitcoin'] | Twitter Web App | False |
| NaN | 😎 BITCOINLIVE is a Dutch platform aimed at inf... | 2019-10-17 20:12:10 | 6769.0 | 1532 | 25483 | False | 2021-02-10 23:58:48 | 😎 Today, that's this #Thursday, we will do a "... | ['Thursday', 'Btc', 'wallet', 'security'] | Twitter for Android | False |
| London, England | IM Academy : The best #forex, #SelfEducation, ... | 2014-11-10 10:50:37 | 128.0 | 332 | 924 | False | 2021-02-10 23:54:48 | Guys evening, I have read this article about B... | NaN | Twitter Web App | False |
| NaN | I will post a lot of buying signals for BTC tr... | 2019-09-28 16:48:12 | 625.0 | 129 | 14 | False | 2021-02-10 23:54:33 | $BTC A big chance in a billion! Price: \487264... | ['Bitcoin', 'FX', 'BTC', 'crypto'] | dlvr.it | False |
| Europa | Co-founder @RENJERJerky \| Forbes 30Under30 \| I... | 2016-02-03 13:15:55 | 1249.0 | 1472 | 10482 | False | 2021-02-10 23:54:06 | This network is secured by 9 508 nodes as of t... | ['BTC'] | Twitter Web App | False |

Fig-14 Raw Data of Twitter

In fig-14, we can see there is a lot of data about 2.2 million tweets and not always it has to be correct or without noise. The noise will always be an issue in a data science project as it always results in outliers or give bad results and give improper models. We need to transform the data to normalise it. These are the files selected for sentiment analysis as shown in fig-15.

```
tweets_raw_file   = 'Bitcoin_tweets.csv'
tweets_clean_file = 'Bitcoin_tweets_clean.csv'
bit_price_file2 = 'BTC-USD.csv'
# bit_price_file2 = 'data/BTC-USD.csv'
```

```
crypto_share_file = 'cryptos_share.xlsx'
```

Fig-15 Files Selected For Analysis

```
df_raw.columns
```

```
Index(['user_name', 'user_location', 'user_description', 'user_created',
       'user_followers', 'user_friends', 'user_favourites', 'user_verified',
       'date', 'text', 'hashtags', 'source', 'is_retweet'],
      dtype='object')
```

Fig-16 Attributes In Twitter Dataset

There are 13 attributes in the selected dataset.

```
df_raw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2259788 entries, 0 to 2259787
Data columns (total 13 columns):
 #   Column            Dtype
---  ------            -----
 0   user_name         object
 1   user_location     object
 2   user_description  object
 3   user_created      object
 4   user_followers    float64
 5   user_friends      object
 6   user_favourites   object
 7   user_verified     object
 8   date              object
 9   text              object
 10  hashtags          object
 11  source            object
 12  is_retweet        object
dtypes: float64(1), object(12)
memory usage: 224.1+ MB
```

Fig-17 Datatypes Of Dataset

Fig-17 shows the data type of each attribute respectively and as we can see the data type of a few attributes is not correct as we are working on time-series analysis the date is an object. We need to change it to datetime format.

```
pd.to_datetime(df_raw['date'], format='%Y-%m-%d %H:%M:%S', errors= 'coerce')
```

```
0         2021-02-10 23:59:04
1         2021-02-10 23:58:48
2         2021-02-10 23:54:48
3         2021-02-10 23:54:33
4         2021-02-10 23:54:06
                  ...
2259783   2022-01-19 19:08:16
2259784   2022-01-19 19:08:08
2259785   2022-01-19 19:08:02
2259786   2022-01-19 19:07:54
2259787   2022-01-19 19:07:52
Name: date, Length: 2259788, dtype: datetime64[ns]
```

Fig-18 Function To Change Dtype

There is a function in pandas, we can use it to change from object to datetime64[ns] as shown in fig-18. This will help in creating graphs and showing visualisations. This function changed the attribute dtype with the format “Year, Month, Date, Hour, Minutes, Seconds” and if there are any errors they are coerced.

Cleaning data is the next process in our project, as there is a lot of noise in the dataset we need to clean it before going to the next step.

```
dd['text'].isnull().values.any()
```

```
False
```

Fig-19 Na Values in Tweet

As we can see there are no Na values in the tweet in fig-19.

**Cleaning Data ¶**

```
# clean df
df_raw = df_raw.sort_values(by = 'date')
dd = df_raw.sample(frac=0.01, replace=False, random_state=1)
dd.reset_index(inplace=True)
for i,s in enumerate(tqdm(dd['text'],position=0, leave=True)):
    text = str(dd.loc[i, 'text'])
    text = text.replace("#", "")
    text = re.sub('https?://(?:[-\w.]|(?:%[\da-fA-F]{2}))+', '', text, flags=re.MULTILINE)
    text = re.sub('@\\w+ *', '', text, flags=re.MULTILINE)
    dd.loc[i, 'text'] = text
# f = open(tweets_clean_file, 'a+', encoding='utf-8')
dd.to_csv(tweets_clean_file, header=True, encoding='utf-8',index=False)
```

Fig-20 Cleaning Data

Fig-20 shows how to clean data by first sorting the data by date and using the random sampling of the dataset with df_raw.sample with a fraction of “0.01” and replacement as “false” and random state is “1” then reset the index. As we know the dataset is having a lot of noise, to solve this issue we use regular expressions to identify the specific type of text and replace them. We used a loop for text attributes in the tweets dataset to identify Hashtags, HTTPS, and any mentions in the tweets that starts with @ special character.

```
dd[['text', 'date']]
```

| | text | date |
|---|---|---|
| 0 | nice project \n\n\n\n\n\nortcoin ort okratech ... | 2021-06-22 05:22:52 |
| 1 | Long Bitcoin short the banks 🤛 | 2021-08-25 07:06:40 |
| 2 | Top Trending Cryptocurrency Post - DOGECOIN Se... | 2021-07-02 19:37:18 |
| 3 | Can one expect another wave of BTC's decline s... | 2021-07-24 14:14:53 |
| 4 | We will see...\n\nbitcoin btc bnb band bake $btc... | 2021-05-29 15:52:17 |
| ... | ... | ... |
| 22593 | Bitcoin is the American dream. Today I have fi... | 2021-07-22 15:18:56 |
| 22594 | If you invested your $1,400.00 stimulus check ... | 2021-08-23 22:00:04 |
| 22595 | when glacier mining\n\n$btc Bitcoin | 2021-06-21 09:09:22 |
| 22596 | ANKR COIN will soon come to its new target of ... | 2021-07-21 20:11:28 |
| 22597 | Today's Titan WOTD: Halving\n\nHalving is when... | 2021-08-25 19:00:00 |

22598 rows × 2 columns

Fig-21 Cleaned Data

In fig-21 we can see through the enumeration we deleted noise and data which is not useful, before the data was with hashtags, https, and special characters.

Another important aspect in EDA is to find the Nan values in 13 attributes of the dataset. It is a quite real-world problem when collecting data from sources it may have occurred due to problems while fetching or blank spaces sometimes. The Nan or null values are commonly denoted as missing data. The issue with the missing data while building a ML is most of the models are not able to process address them. There are few ways to deal with missing data in the attributes, if it is a numerical data type we can find the mean of the variable and update missing information and if it is a categorical data we can either delete the row or multiple rows.

```
for miss_dd in dd:
    print({miss_dd: dd[miss_dd].isnull().sum().sum()})
```

```
{'index': 0}
{'user_name': 0}
{'user_location': 11056}
{'user_description': 2705}
{'user_created': 1}
{'user_followers': 2}
{'user_friends': 2}
{'user_favourites': 2}
{'user_verified': 2}
{'date': 2}
{'text': 0}
{'hashtags': 178}
{'source': 33}
{'is_retweet': 4}
```

Fig-22 Missing Values In Each Attribute

We need to check for Na values because we need to select the meaningful features as we can see later, having redundancy in our data leads Machine Learning models to overfit causing the model to be overly complicated for training and unable to find generalize well on unseen data. One of the ways to reduce the overfitting is to lower the dimensionality. This is often done by principal component analysis (PCA) a type of unsupervised machine learning to reduce the number of features in a dataset. Based on different patterns in characteristics, PCA finds relationships in our data.

To find how many missing values are in the dataframe we can check with following function in pandas as shown in fig-22. There are few columns where there are no missing values where as user_location has 11056 items, user_description have 2705 items, user_created have 1, user_followers have 2, user_friends have 2, user_favourites have 2 , date have 2, hashtags have 178, source have 33, is_retweet have 4.

This correlation suggests that there is a redundancy in our data or there are some parts of it may be explained. The non-correlated data is not required for model to learn its weights and may be removed. It can be deleted either directly by columns (features) or by merging in some cases through the process to save the important part of information. There are only few columns which we need to check for minimum outliers or missing data, as we can see the “text” column have 0 missing values in data and there is no need to perform normalization or standardization.

## Building Model

We are building a model based on Vader and LSTM models, where Vader is a lexicon-based approach where it will take into consideration one of two main approaches to sentiment and involves in calculating the sentiment score by semantic orientation of words or phrases occurring in a text. A rule-based analysis sentiment tool that is specific accustomed to expressions in social media. It considers negations and contractions like good, not good, punctuations (!), emotes, caps, acronyms (lol) scores between -1 as negative, 1 as positive. We will use this score method to calculate tweets score that will represent importance of each tweet.

To generate scores in dataset we will use “vadersentiment” module in python. Using the cleaned tweets dataset, we are ready to give score for each text column. Using analyzer from sentiment intensity analyzer we will give polarity scores by creating a new column called

compound. The compound column represents the tweets sentiment by its value between -1 and 1. We will use number of tweets, number of likes and number of users that will follow the user.

```
analyzer = SentimentIntensityAnalyzer()
compound = []
for i,s in enumerate(tqdm(df_clean['text'],position=0, leave=True)):
    # print(i,s)
    vs = analyzer.polarity_scores(str(s))
    compound.append(vs["compound"])
df_clean["compound"] = compound
df_clean.head(5)d
```

```
100%|                                                                                | 22598/22598 [00:02<00:00, 10563.21
it/s]
```

| description | user_created | user_followers | user_friends | user_favourites | user_verified | date | text | hashtags | source | is_retweet | compound |
|---|---|---|---|---|---|---|---|---|---|---|---|
| خوبت رو به خو نیگران | 2021-01-02 19:08:56 | 616.0 | 1683.0 | 1712.0 | False | 2021-06-22 05:22:52 | nice project \n\n\n\n\n\nortcoin ort okratech ... | ['ortcoin', 'ort', 'okratech', 'bitcoin', 'air... | Twitter for Android | False | 0.5994 |
| volution the nge 👊😎👊 Not y... | 2019-10-13 06:16:05 | 451.0 | 2045.0 | 45077.0 | False | 2021-08-25 07:06:40 | Long Bitcoin short the banks 👏 | ['Bitcoin'] | Twitter Web App | False | 0.0000 |
| igital Token r | Top Mod @ https:/... | 2014-04-25 19:40:15 | 468.0 | 111.0 | 1262.0 | False | 2021-07-02 19:37:18 | Top Trending Cryptocurrency Post - DOGECOIN Se... | ['Cryptocurrency', 'Bitcoin'] | IFTTT | False | 0.2023 |
| 6MRU7xb1 dent orga... | 2018-01-17 09:28:59 | 35990.0 | 4447.0 | 682.0 | False | 2021-07-24 14:14:53 | Can one expect another wave of BTC's decline s... | ['BTC'] | get shares | False | 0.0000 |
| alized, be a art contract | 2010-08-23 20:41:38 | 164.0 | 100.0 | 9295.0 | False | 2021-05-29 15:52:17 | We will see... \n\nbitcoin btc bnb band bake $btc... | ['bitcoin', 'btc', 'bnb', 'band', 'bake'] | Twitter for Mac | False | 0.0000 |

Fig-23 Sentiment Score

We do not need all the columns in the dataset to perform analysis we can drop them, we need columns which may have higher correlation ‘user_followers’, ‘user_friends’, ‘user_favourites’, ‘compound’, ‘score’.

## Pearson-Cross Correlation

It is a process of identifying a link between two time series and variables. It is the metric for determining a linear relationship of a number which ranges from 1 to -1 that shows the direction of the link between two variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r - correlation coefficient

$x_i$ - values of x-variable in sample

$\bar{x}$ - mean of the values in x-variable

$y_i$ - values of y-variable in sample

$\bar{y}$ - mean of the values in y-variable

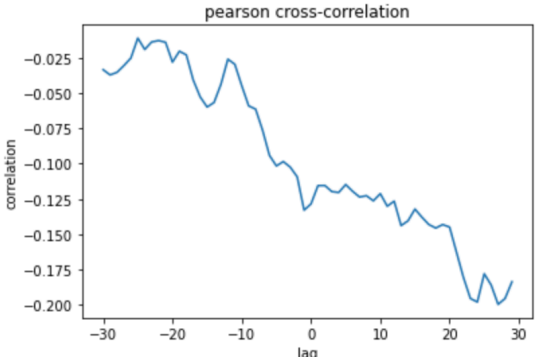


Fig-24 Pearson cross-correlation

| Correlation value | Type | Meaning |
|---|---|---|
| +1 | Positive correlation | When one variable changes, other will change in same direction |
| 0 | Neutral | No relation between variables |
| -1 | Negative | When one variable changes, other will change in opposite direction |

## Kendall Cross-Correlation

It is a correlation that measures the degree and direction of link between two variables (determines if a monotonic relationship exists).

τ = (C-D) /(C+D)
C – number of concordant pairs
D – number of discordant pairs

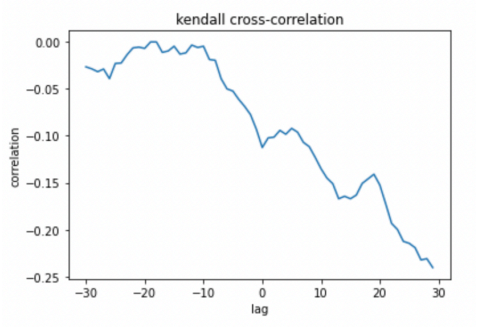


Fig-25 Kendall Cross-Correlation

## Spearman's Correlation

It is the cross-correlation that measures the degree and direction of monotonicity is "less constraining" than linear connection between variables.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$\rho$ - spearman's rank correlation
$d_i$ - difference of two ranks of each observation
$\eta$ - number of observations

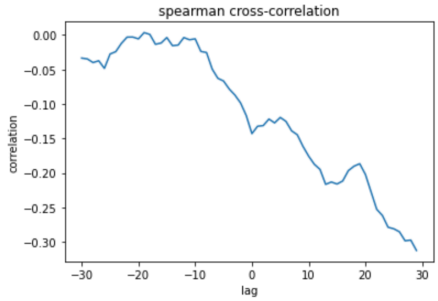


Fig-26 Spearman's correlation

We can see from all three types of correlation that both Spearman and Kendall are the same, but Pearson correlation graph is significantly different due to data not being normalized. It is important to have the data to be normalized to remove outliers and give good results.

## Normalization

Normalization is the pre check in the data science project, where it is a common technique of converting the columns in the dataset to the same scale. The mostly used types of scaling are min-max scaling and standardization scaling, where min-max is the minimum value of each column's highest value divided by range of each column has minimum value of 0 and a maximum value of 1 and term standardization is the process of centring a variable at 0 and standard variance at 1.

In our project we will be using min-max technique to normalize the data to transform the dataset to fit tweets which are grouped and assigned to a new column as normalized_score. Then the grouped crypto dataset is divided with maximum of highest grouped crypto value and minimum grouped crypto. Below is the graph fig-27 plotted after normalizing the data of bitcoin and sentiment.

Scaled group of crypto = Grouped crypto / max(grouped crypto maximum value*grouped minimum value)

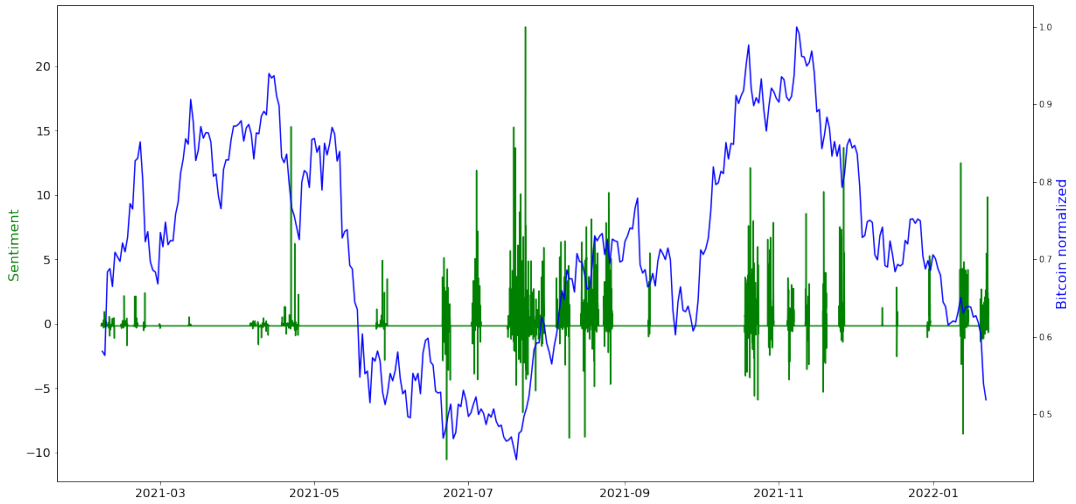


Fig-27 Normalized crypto currency evolution compared to twitter data

Here we can see the tweets are few patterns in graphs that indicate from July-2021 to Sept-2021 there was a highest sentiment in twitter and after normalizing the bitcoin price we can see that the price of bitcoin was raising of its value in that time period or people are reacting to the price more at the same time. It might be a relation that shows people are reacting positively when bitcoin was raising that indirectly pushes the bitcoin price to go up and vice versa, when bitcoin price falls people might be reacting negatively that in turn results in more reduction of its original value.

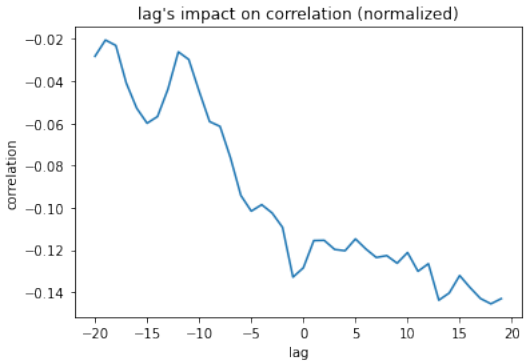


Fig-28 Correlation of Normalized Data

Fig-28 shows the correlation of lag's impact of tweets and crypto which are grouped together.

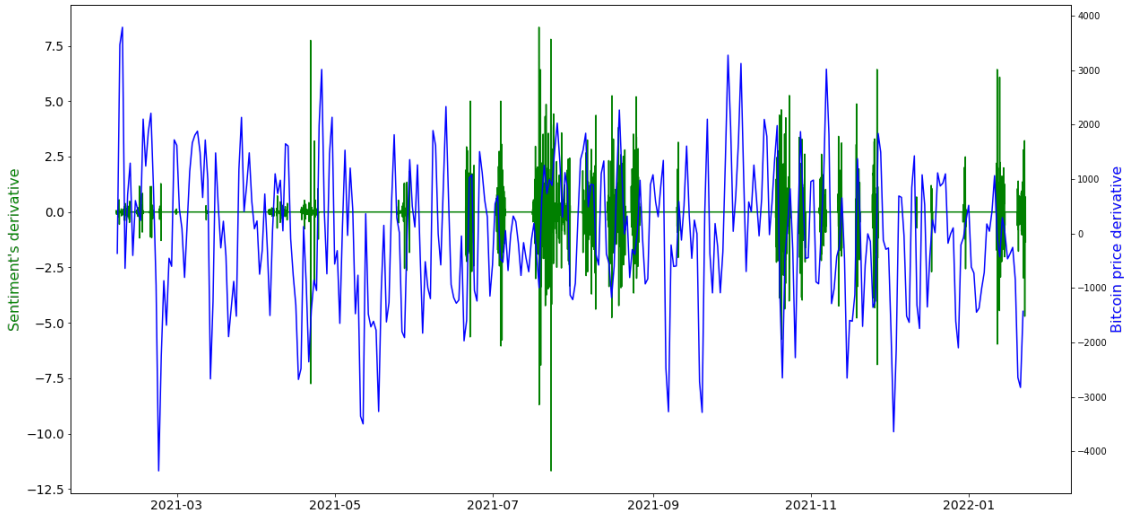


Fig-29 Derivative of Crypto Price

The term derivate in financial aspect is the contract that derives its price from the performance of underlying entity which might be index fund, asset, interest rate and is often called as underlying which perfectly shows in fig-29 of derivate of bitcoin price and sentiment score. Now if we check the cross correlation of Pearson, Kendall, Spearman we can see the correct relation between tweets and bitcoin price.

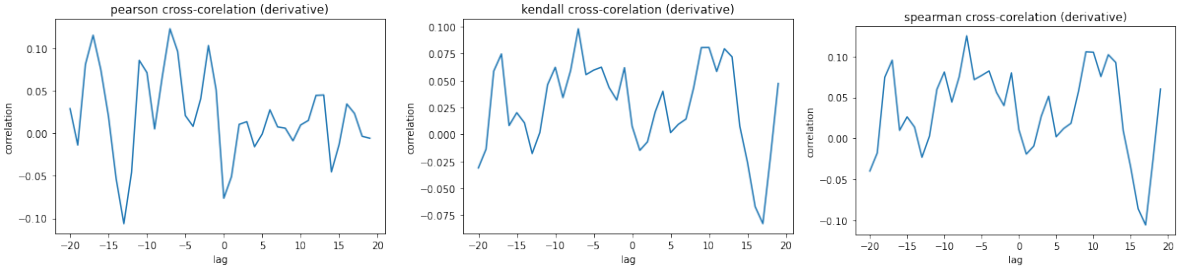


Fig-30 Derivative of Correlations

The graphs indicate there is normalized correlation between bitcoin price and twitter users reacting to its price.

## NLP Modelling

NLP modelling is the technique of finding the patterns in a text that specifies formats of words and output of excellence. NLP modelling is the first activity while working on text analysis, and the process extends to NLP applications. These are typically taking the form of real-valued vectors of words, these word embeddings are generated by employing a collecting of language modelling and feature learning algorithms in which lexicon-based mapping to real-number vectors.

We will use word lemmatizing by doing a vocabulary and morphological analysis of words. Using module from python nltk we can download stop words like 'i', 'me', 'we', 'our', etc., finding these words we can eliminate them in analysis as these kinds of words will not have impact on positive or negative statements.

```
#remove urls
tweet_without_url = re.sub(r'http\S+',' ', data)

#remove hashtags
tweet_without_hashtag = re.sub(r'#\w+', ' ', tweet_without_url)

#3. Remove mentions and characters that not in the English alphabets
tweet_without_mentions = re.sub(r'@\w+',' ', tweet_without_hashtag)
precleaned_tweet = re.sub('[^A-Za-z]+', ' ', tweet_without_mentions)

#2. Tokenize
tweet_tokens = TweetTokenizer().tokenize(precleaned_tweet)

#3. Remove Puncs
tokens_without_punc = [w for w in tweet_tokens if w.isalpha()]

#4. Removing Stopwords
tokens_without_sw = [t for t in tokens_without_punc if t not in stop_words]

#5. lemma
text_cleaned = [lem.lemmatize(t) for t in tokens_without_sw]
```

Fig-31 Cleaning Dataset and Tokenizing

We should clean the tweets column with tweet tokenizer and lemmatize by removing specific kinds of words like above, https links, hashtags, and whitespaces. Remove mentions in a tweet which are basically names of people and special characters like symbols that will not have any sentiment then tokenizing the words in the columns. After we do this process we will get cleaned columns column as below.

| | tweets | cleaned_tweets | date_clean |
|---|---|---|---|
| 0 | nice project \n\n\n\n\n\nortcoin ort okratech ... | nice project ortcoin ort okratech bitcoin aird... | 2021-06-22 |
| 1 | Long Bitcoin short the banks ✊ | Long Bitcoin short bank | 2021-08-25 |
| 2 | Top Trending Cryptocurrency Post - DOGECOIN Se... | Top Trending Cryptocurrency Post DOGECOIN Sell... | 2021-07-02 |
| 3 | Can one expect another wave of BTC's decline s... | Can one expect another wave BTC decline soon v... | 2021-07-24 |
| 4 | We will see…\n\nbitcoin btc bnb band bake $btc... | We see bitcoin btc bnb band bake btc dCc xZ dP | 2021-05-29 |

Another important process in our tweet analysis is finding polarity and subjectivity of our tweets text.

```
def getSubjectivity(tweet):
    return TextBlob(tweet).sentiment.subjectivity

def getPolarity(tweet):
    return TextBlob(tweet).sentiment.polarity
```

Polarity is the strength of an opinion in a text, it might be both good and bad. If anything is linked with strong connection of good emotion like love, trust, like will have orientation of all the parts of the object's existence, this will be same for negative polarities. Positive and Negative polarity can change in strength depending on the context, sentiment analysis techniques would be classified as either slightly positive sentiment or negative sentiment. What interests most is how people will express in nonverbal way.

Subjectivity is the degree of person is connected with any item. Individual and Personal connections are with an item that are important which may or may not defer from someone's point of view.

Sentiment analysis algorithms must be able to discern with the polarity or subjectivity to examine users' thoughts when conveyed. A lower degree may represent someone point of view on a topic. Sentiments can also have varying degree of polarity across different methods of communication in emails, chats, messages which might be positive, negative, or neutral.

```
def crypto_price_cate(score):
    if score < 1:
        return 'negative'
    elif score == 1:
        return 'neutral'
    else:
        return 'positive'

def observe_period(period):
    res = crypto_usd_grouped.shift(period)/crypto_usd_grouped
    res = res.apply(crypto_price_cate)
    return res

time_sentiment = observe_period(7) # compare price ratio in 7 days. price_7_days_later/ price_now
df['crypto_sentiment'] = df.date_clean.apply(lambda x: time_sentiment[x] if x in time_sentiment else np.nan)
```

Fig-32 Score Assignment

This function allows us to find the negative, neutral, or positive sentiment towards the bitcoin tweets. Where if the score is less than 1 it is considered as negative sentiment, if the score is equal to 1 it is neutral sentiment and if it is more than 1 it is positive sentiment.

| | tweets | cleaned_tweets | date_clean | crypto_sentiment | subjectivity | polarity |
|---|---|---|---|---|---|---|
| 0 | nice project \n\n\n\n\n\nortcoin ort okratech ... | nice project ortcoin ort okratech bitcoin aird... | 2021-06-22 | positive | 1.00 | 0.600 |
| 1 | Long Bitcoin short the banks 😠 | Long Bitcoin short bank | 2021-08-25 | negative | 0.35 | -0.025 |
| 2 | Top Trending Cryptocurrency Post - DOGECOIN Se... | Top Trending Cryptocurrency Post DOGECOIN Sell... | 2021-07-02 | negative | 0.40 | 0.250 |
| 3 | Can one expect another wave of BTC's decline s... | Can one expect another wave BTC decline soon v... | 2021-07-24 | negative | 0.00 | 0.000 |
| 4 | We will see...\n\nbitcoin btc bnb band bake $btc... | We see bitcoin btc bnb band bake btc dCc xZ dP | 2021-05-29 | positive | 0.00 | 0.000 |

Fig-33 Finding Subjectivity and Polarity

The above figure shows sentiment of subjectivity and polarity in tweets scores. Now the dataset is ready for model, as we are going to build it we need to split data into training and testing split as discussed in methodology. Training set should be large enough to yield statistically meaningful results and we should be aware that we create variables according to that as we cannot build model on test variable.

# 7. Training:

The random state is put to 38 so that how many times we may run the code the model results will be same in train and test values and the random numbers would be generated will be constant.

```
seed = 38 # fix random seed for reproducibility
np.random.seed(seed)

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2,
                                                    stratify=y,
                                                    random_state=seed)
```

Fig-34 Splitting Train and Test Data

We use sequential, dense, embedding, conv1d, maxpooling and LSTM models from Sklearn to create a model. There are 7 layers, which have total of 2,066,239 total params and trainable params by distributing vectors.

**Model-1**

Model: "sequential"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding (Embedding) | (None, 30, 100) | 2000000 |
| conv1d (Conv1D) | (None, 30, 32) | 9632 |
| max_pooling1d (MaxPooling1D) | (None, 15, 32) | 0 |
| conv1d_1 (Conv1D) | (None, 15, 32) | 3104 |
| max_pooling1d_1 (MaxPooling1D) | (None, 7, 32) | 0 |
| lstm (LSTM) | (None, 100) | 53200 |
| dense (Dense) | (None, 3) | 303 |

Total params: 2,066,239
Trainable params: 2,066,239
Non-trainable params: 0

Relu is the function used most in the machine learning projects as it works on linear activation. It will give us the state of art results computationally and very efficient at the same time.

$$f(x) = \max(0, x) = \begin{cases} x_i \; if \; x_i > 0 \\ 0 \; if \; x_i < 0 \end{cases}$$

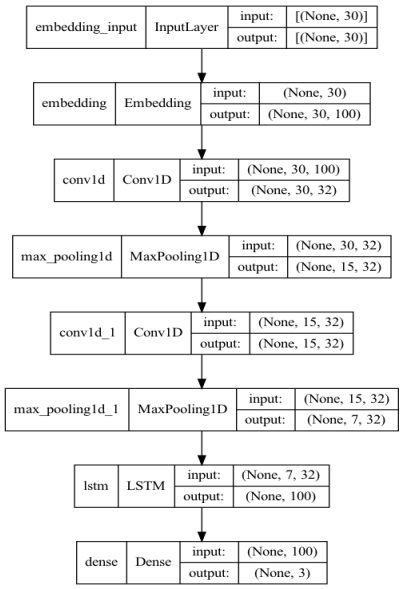


Fig-35 Model-1

Fig-35 shows the total number of layers we used to predict or depict the results including hidden layers adding to the output, this model is single directional learning.

**Model-2**

Model-2 is a bidirectional layer inputs where the transfer rate of NLP model is both ways to check which model shows the better results.

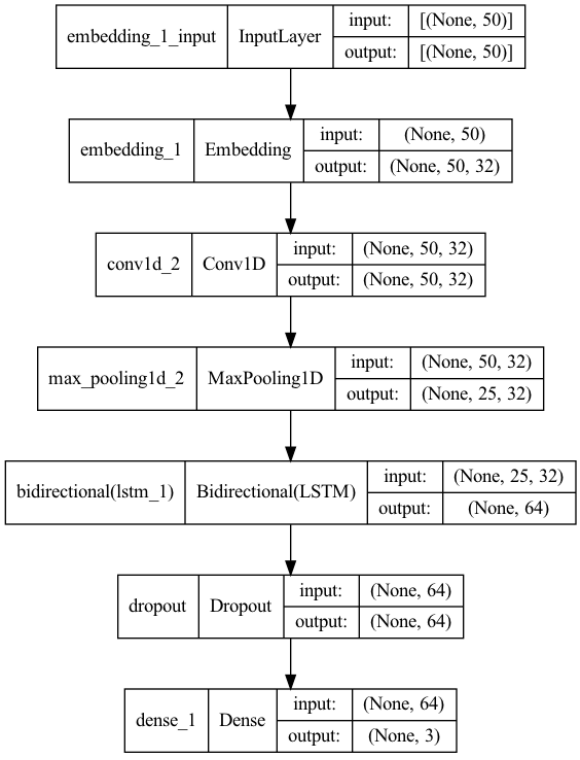


Fig-36 Model-2

The lower the loss of a model the better the accuracy we get from the model “H. et al., (2015)”, unless the model has an over-fitting data. The loss is measured on training and validation and its interpretation how well a model is built.

The accuracy of a model is typically measured after the model parameters are learned and there is no longer learning required. The test samples are then fed into model and then the numbers of errors made by the model are recorded comparing to the target as seen in fig-37. The accuracy of the model as seen in the graph is above 90 %.

| | tweets | cleaned_tweets | date_clean | crypto_sentiment | subjectivity | polarity | sentiment | target |
|---|---|---|---|---|---|---|---|---|
| 0 | nice project \n\n\n\n\n\nortcoin ort okratech ... | nice project ortcoin ort okratech bitcoin aird... | 2021-06-22 | positive | 1.00 | 0.600 | positive | True |
| 1 | Long Bitcoin short the banks 😉 | Long Bitcoin short bank | 2021-08-25 | negative | 0.35 | -0.025 | negative | True |
| 2 | Top Trending Cryptocurrency Post - DOGECOIN Se... | Top Trending Cryptocurrency Post DOGECOIN Sell... | 2021-07-02 | negative | 0.40 | 0.250 | positive | False |
| 3 | Can one expect another wave of BTC's decline s... | Can one expect another wave BTC decline soon v... | 2021-07-24 | negative | 0.00 | 0.000 | neutral | False |
| 4 | We will see...\n\nbitcoin btc bnb band bake $btc... | We see bitcoin btc bnb band bake btc dCc xZ dP | 2021-05-29 | positive | 0.00 | 0.000 | neutral | False |

Fig-37 Target Column With Accuracy Model vs loss

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.67 | 0.76 | 478 |
| 1 | 0.94 | 0.94 | 0.94 | 1782 |
| 2 | 0.93 | 0.97 | 0.95 | 2224 |
| accuracy | | | 0.93 | 4484 |
| macro avg | 0.91 | 0.86 | 0.88 | 4484 |
| weighted avg | 0.93 | 0.93 | 0.92 | 4484 |

# 8. Testing:

The accuracy is metric of a classification model that quantifies as the number of correct predictions by the number of total predictions. The model obtains high f-1 score if the precision and recall are more, model obtains medium if these are neither low or high, and the model obtains low f-1 score if these are low. As from above table we can see the model accuracy is 93% and as we can see f-1 score is increased from 76% to 95% while modelling support from precision 0.93, 0.97 and 0.87, 0.67 respectively.

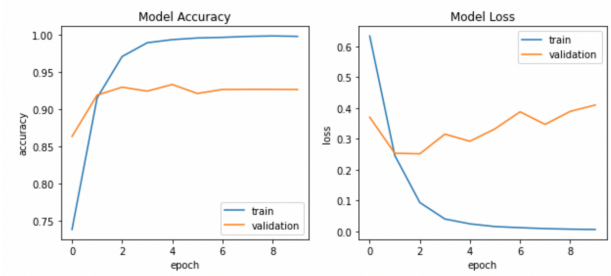


Fig-36 Model-1 Accuracy & Loss

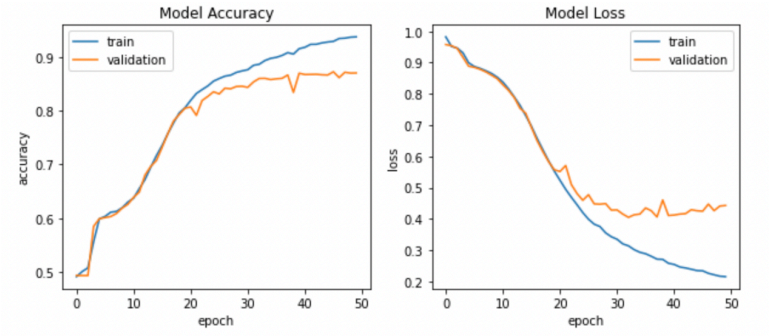


Fig-38 Model-2 Accuracy & Loss

Model-2 accuracy shows that validation data is not so accurate with the first model. The Accuracy of model-2 is at 87%, Precision is at 89%, Recall at 86% and F1 Score is at 87.

**Confusion Matrix**

After cleaning, wrangling and data pre-process, the first process is to feed it to the exceptional model and output in probabilities. In order to measure the effectiveness of our models we can use confusion matrix to get a clear picture which is a performance measurement for ML classification. The confusion matrix is a table layout that indicates how well the performance of an algorithm is working. Each row of the matrix represents an actual class, whereas each column represents a projected class or vice-versa, it is a simple visualization that explains whether a system is confusing two types of classes.

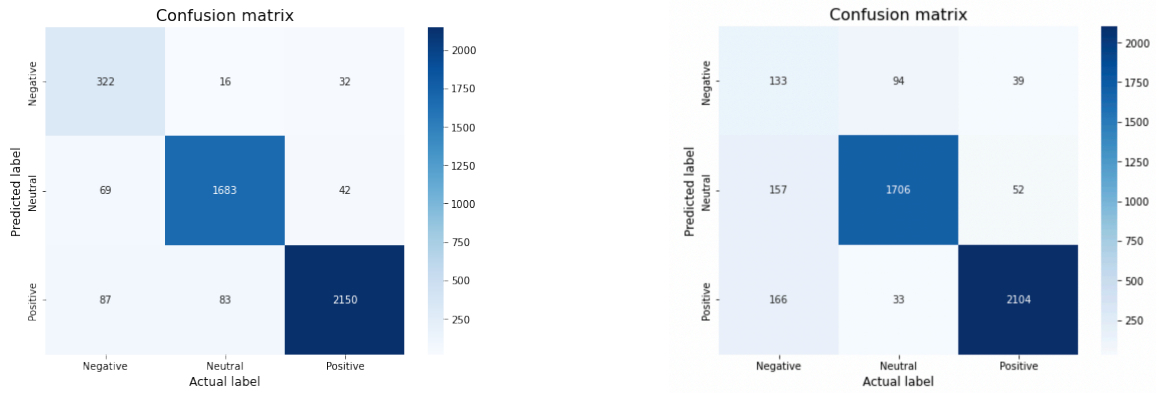


Fig-39 Confusion Matrix of Model-1 & Model-2

| | | Condition | |
|---|---|---|---|
| | Total population = P+N | Positive (PP) | Negative(PN) |
| Actual Condition | Positive(P) | True Positive (TP) | False Negative (FN) |
| | Negative(N) | False Positive (FP) | True Negative (TN) |

The above table represents the positive and negative effect on the tweets by users. Fig-39 shows the true aspect of people which type of sentiment from the model.

# 9. Project Management:

Every project needs to have a plan and management to follow certain methodologies, debugging, and a result-oriented approach. This section mainly focusses on the how well a project is maintained, effective according to rota from the proposal. We need to take into consideration of time, code of ethics, planning, risk assessment, maintaining a correct timetable so that we will meet the output result for the project. As an outcome of the risk-assessment, development through analysis, and creating a model we need to follow few steps to project deliverables.

As we are following agile methodology, project management should be maintained according to the planned rota and if there are any new updates in the process we should be able to adapt to the new plan. This methodology will ensure the project is completed on time with good quality of project. The process included stages like problem statement, requirements, design, coding and development, integration and testing models, and review.

The key component of project management includes project timeline, quality management, risk management, social , legal, ethical, and professional issues as discussed below.

### Project Schedule

The project took 9 weeks to complete along with report. There were few steps to start the project as this is a resit submission I chose bitcoin sentimental analysis to do due to the tight timeline. The project dealt with finding the emotion of people towards the price changes on how twitter users are reacting to the fluctuations. This schedule involved in 3 steps creating ethical document, writing proposal, and report submission, following table will show the information of date timeline.

| **Task Deliverables** | **Date** |
|---|---|
| Ethics Document | 09-06-2022 |
| Proposal | 24-06-2022 |
| Report | 01-08-2022 |

This table is the basic submission and deadlines of the project, but when it comes to project management I followed few steps as follows:

Task-1: Topic selection
Task-2: Project plan, Collecting Data
Task-3: Ethics form submission
Task-4: Proposal, Literature Review
Task-5: Data pre-process & Analysis
Task-6: Building Model, Writing report
Task-7: Evaluating Models
Task-8: Finishing Report

| Task | W1 | W2 | W3 | W4 | W5 | W6 | W7 | W8 |
|---|---|---|---|---|---|---|---|---|
| Task-1 | | | | | | | | |
| Task-2 | | | | | | | | |
| Task-3 | | | | | | | | |
| Task-4 | | | | | | | | |
| Task-5 | | | | | | | | |
| Task-6 | | | | | | | | |
| Task-7 | | | | | | | | |
| Task-8 | | | | | | | | |

Every fortnight on Wednesday I had attended the calls with prof. simon billings to discuss about project, got guidance in building models, completing tasks above and creating this report.

## Risk Management

There will be some risk in maintaining any project, building code and report writing as the data can be lost due to laptop damages or such. After analysing overall risk in the data breach or data loss with the implementation, dangers, and risk the solution is as follows. Some kinds of risks that follows in project are:

1. Data loss – there is good amount of chance that any kind of data will loose from any device when storing locally. To avoid this problem, we need to use cloud services, and to avoid data breach from public servers it is always recommended to use trusted sources provided from university.
2. Hardware issues – hardware is the main component in data science project, where we store code, data, visualizations etc., if there is an issue with the hardware it will be difficult to meet requirements at the deadline submission for code. It is recommended to store code, and data in repositories.
3. Memory – bigdata can be another issue during program executions. If there is less memory for building models it will take most of our coding time to just run it and sit back. It is recommended to have good hardware and memory. If there is an issue with GPU on local computers we can borrow a good computer or use Google colab.

## Quality Management

This section focuses on the quality of the project implementation. Firstly, good programming language to be selected to perform analysis and modelling, python is chosen to download analysis modules and machine learning models.

Attending supervisor meetings is another way of maintaining quality of the project. Teams meeting with fellow colleagues and supervisor fortnight helps in suggestions, doubt sessions and recommendation on improving the standards of analysis. During these meeting an update of pending items and goal for the next week will be made on appendix-A.

### Social , Legal, Ethical and Professional Considerations

There are few issues and concerns to be addressed before the project can be implemented. To continue with the project, we need ethics approval from Coventry university because any data-based projects have concerns raising for protected data or conducting a research of data breaches or privacy of an individual is exposed. As this is a sensitive topic it is important to have appropriate permissions with the university to get started and take user privacy privileges seriously.

To demonstrate this, a questionnaire is filled in https://ethics.coventry.ac.uk/ website and submitted to the supervisor. The application then will be reviewed by supervisor and approved if there are no issues with the project. Later at this point we can start the project. Ethics approval form is submitted in appendix-B.

# 10. Critical Appraisal:

The aim of this project is to study or research and create machine learning models that will find sentiment on bitcoin tweets. Traditional machine learning models are used in the research, the model predicted were improved in this research after doing literature review from the old research methods but are unrealistic due to time-series and always changing data. Many among the others will use neural network models when dealing with the time-series to execute realistic rather than simple machine learning models. I tested with two models to do research to check how well they are performing. Vader sentiment with LSTM model gave good results.

Many of the modules I studied in this course helped me understanding about data science role in day-to-day life. This research helped me to understand about how emotions can impact in finding how people are reacting to a new technology called bitcoin. This project can be used find critical analysis of emotions of tweets about a topic, technology, crypto etc., by making few tweaks in the code.

# 11. Conclusions:

The time-series models are analysed used by python programming language and pre-processing the data to make it ready for models and pass it through neural network models in finding emotions of people with tweets. The goal of this research is to develop a better machine learning model for predicting sentiment on bitcoin prices. LSTM and different kind of visualisations are used in the code to demonstrate my understanding of the data. Using machine learning models for time-series data we can see how accurate they are forecasting, even if the model is performing in unrealistic volume. In terms of time-series data prediction, neural network models are significantly more realistic.

According to the research I conducted by comparing two models with the f-1 scores, precision and recall it is clear that 1$^{st}$ model is clearly performing better finding the score with 93% and 2$^{nd}$ model is performing with a little low accuracy of 87%. Predictions for near values are shown in the findings are described with plots, and graphs. This study helped me in understanding extensively on machine learning models, neural networks, and sentiment.

Using this model an investor will have idea, how the price can shift and he/she can decide when to buy or when to sell bitcoin.

### Achievements

Any models can be improved in a better way, although models performing on time-series datasets might be good or poor. The project goal might be realised at the ending of the execution and can be used in depth.

### Future Work

Although project could be produced in a better way there are many concepts that aren't done due to time for the implementation and deadline.

1. Improving gathering data of tweets and bitcoin prices live by using API's.
2. Performance tuning by adding day-to-day.
3. This project can be used for any other crypto price sentiment or stock market sentiment etc.,
4. Automating gathering data using tools like Airflow.
5. Long-term or short-term analysis can be tried to implement model for intervals of time.
6. Data used is mainly from twitter, we can try to use more modes of user's sentiment by gathering them from different sources.
7. We have mostly worked on English language tweets, models can also be trained on different languages finding their patterns.

# 12. Student Reflections:

This research gave me chance of understanding machine learning concepts well for my career. Through the course of master's degree, I was able to acquire good knowledge and support from professors even though I came from non-technical background from civil engineering from bachelors. This was vital because being a core engineer inspired by the impact of analysis in the world I did learn more than ever in this year. The first project submitted by me has few issues because I had covid at that time and due to health reasons I couldn't manage time well but getting another chance to create a new project and submit this research paper. The last few weeks were difficult during the assignment were difficult because of the pressure and financial situations I was facing. Aside from that I managed to learn everything and applied my mind and heart into this project theoretically.

My professor Simon billings supported me in every aspect he could in this difficult time, guided me to create good project and submit it.

Twitter sentiment prediction helped me in learning new analysis strategy, showing sentiment also is an investment and you can get profits by studying them. I began learning about this process from scratch and had a chance of reading some good books, literature papers, understanding and implementing them to the data. Any price fluctuations like bitcoin, stocks or bonds will always move with run with peoples' emotions. Overall, I was able to get a good

knowledge on a new business model and I can use this knowledge to implement when I am working on a business or my future self.

## References:


1. Wikipedia contributors. (2022, July 3). Bitcoin. Wikipedia. https://en.wikipedia.org/wiki/Bitcoin#cite_note-NY2011-17
2. Bloomberg - (2022). Retrieved 12 July 2022, from https://www.bloomberg.com/news/articles/2022-04-05/ukraine-readies-nft-sales-as-crypto-donations-top-60-million#:~:text=Ukraine%20had%2032%20Bitcoins%20valued,and%20medical%20supplies%2C%20Bornyakov%20said.
3. Moné, L. (2022). The Difference Between Artificial Intelligence, Machine Learning, and Deep Learning. Retrieved 12 July 2022, from https://www.leanix.net/en/blog/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning
4. 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*. doi: 10.1007/s00521-020-05129-6
5. Using sentiment analysis to predict interday Bitcoin price movements | Emerald Insight. (2022). *The Journal Of Risk Finance*, *19*(1), 56-75. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/JRF-06-2017-0092/full/pdf?title=using-sentiment-analysis-to-predict-interday-bitcoin-price-movements
6. Sentiment Analysis: Concept, Analysis and Applications. (2018). Retrieved 10 July 2022, from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
7. Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis. (2020, November 4). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9351527/;jsessionid=ietzsfngSFOKoX3MQy_PMriPYyilt76ZfNbE7tHi-5Vy1G__wdCu!-508165331?tp=&arnumber=9351527
8. Raviv, P. (2018, June 9). Blockchain Makes Sentiment Analysis Made Affordable to All. Bitcoinist.Com.https://bitcoinist.com/blockchain-makes-sentiment-analysis-made-affordable/
9. Theseus.fi. 2022. The impact of Twitter user sentiment on Bitcoin pricing value. [online] Available at: <https://www.theseus.fi/bitstream/handle/10024/510978/Final%20Thesis%20Report.pdf?sequence=2> [Accessed 17 June 2022].
10. Xu, J., Bai, W., Hu, M., Tian, H., & Wu, D. (2020). "Bitcoin miners: Exploring a covert community in the Bitcoin ecosystem." (“Bitcoin miners: Exploring a covert community in the Bitcoin ecosystem ...”) (“Sci-Hub | Bitcoin miners: Exploring a covert community in the Bitcoin ...”) Peer-to-Peer Networking and Applications, 14(2), 644–654. https://doi.org/10.1007/s12083-020-01021-1
11. S. E. Freeda, T. C. E. Selvan and I. G. Hemanandhini, "Prediction of Bitcoin Price using DeepLearning Model," 2021 5th International Conference on Electronics,

Communication and Aerospace Technology (ICECA), 2021, pp. 1702-17DOI doi: 10.1109/ICECA52323.2021.9676048
12. Georgoula et al., 2015 Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D.N., Giaglis, G.M., 2015. "Using time-series and sentiment analysis to detect the determinants of bitcoin prices." ("(PDF) Using Time-Series and Sentiment Analysis to Detect the ...") SSRN Electron. J.
13. Garcia and Schweitzer, 2015 D. Garcia, F. Schweitzer Social signals, and algorithmic trading of Bitcoin Roy. Soc. Open Sci., 2 (9) (2015), p. 150288
14. Mai et al., 2015 F. Mai, Q. Bai, Z. Shan, X.S. Wang, R.H. Chiang From Bitcoin to big coin: the impacts of social media on Bitcoin performance SSRN Electron. J. (2015), pp. 1-16
15. Li et al., 2014 X. Li, H. Xie, L. Chen, J. Wang, X. Deng News impact on stock price return via sentiment analysis Knowl.-Based Syst., 69 (1) (2014), pp. 14-23
16. Karalevicius et al., 2018 V. Karalevicius, N. Degrande, J. De Weerdt Using sentiment analysis to predict interday Bitcoin price movements J. Risk Financ., 19 (1) (2018), pp. 56-75
17. 4 Key Concepts of Bitcoin - GeeksforGeeks. (2021). Retrieved 12 July 2022, from https://www.geeksforgeeks.org/4-key-concepts-of-bitcoin/
18. The economics of Bitcoin price formation. (2022). *Applied Economics*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/00036846.2015.1109038
19. Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal Of International Financial Markets, Institutions And Money*, *65*, 101188. DOI: 10.1016/j.intfin.2020.101188
20. V. Pagolu, K. Challa, G. Panda, Sentiment Analysis of Twitter Data for Predicting Stock Market Movements, 2016 https://arxiv.org/pdf/1610.09225.pdf.
21. M. Matta, I. Lunesu, M. Marchesi, "Bitcoin Spread Prediction Using Social and Web Search Media". Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management – 2015.
22. E. Stenqvist, J. Lönnö, "Predicting Bitcoin price fluctuation with Twitter sentiment analysis" (2017).
23. Bloomberg - Are you a robot?. (2022). Retrieved 14 July 2022, from https://www.bloomberg.com/opinion/articles/2021-05-17/elon-musk-controls-bitcoin-and-dogecoin-prices-with-pure-magic
24. Bambrough, B. (2022). https://www.forbes.com/sites/billybambrough/2022/06/16/elon-musk-suddenly-hit-with-huge-258-billion-dogecoin-pyramid-scheme-lawsuit-amid-a-devastating-bitcoin-and-ethereum-price-crash/
25. F. A. d. Oliveira, L. E. Zarate, M. d. A. Reis and C. N. Nobre, "The use of artificial neural networks in the analysis and prediction of stock prices," in IEEE International Conference on Systems, Man, and Cybernetics, 2011.
26. D. Shah and K. Zhang, "Bayesian regression and Bitcoin," in 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015.
27. Data Preprocessing in Machine Learning [Steps & Techniques]. (2022). Retrieved 15 July 2022, from https://www.v7labs.com/blog/data-preprocessing-guide
28. Python | Sentiment Analysis using VADER - GeeksforGeeks. (2019). Retrieved 15 July 2022, from https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
29. SENTIMENTAL ANALYSIS USING VADER. (2020). Retrieved 19 July 2022, from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

30. NLP | Sentiment Analysis using LSTM - Analytics Vidhya. (2021). Retrieved 19 July 2022, from https://www.analyticsvidhya.com/blog/2021/06/natural-language-processing-sentiment-analysis-using-lstm/
31. H. (2015). How to interpret loss and accuracy for a machine learning model. Retrieved 28 July 2022, from https://stackoverflow.com/questions/34518656/how-to-interpret-loss-and-accuracy-for-a-machine-learning-model

# Appendix A – Interim Progress Report and Meeting Records

| | Meeting Date | Discussion |
|---|---|---|
| 1 | 01-06-2022 | Selection of topic, ethics application, project proposal, dataset |
| 2 | 15-06-2022 | Ethics feedback, Literature Review, weekly updates on EDA |
| 3 | 29-06-2022 | Implementing EDA, plotting, Data pre-process |
| 4 | 13-07-2022 | Modelling, Writing reports |
| 5 | 27-07-2022 | Report discussion and submission |

# Appendix B – Certificate of Ethics Approval



## Certificate of Ethical Approval

| | |
|---|---|
| Applicant: | Vamsidhar Muchurami |
| Project Title: | LEVERAGING TWITTER SENTIMENT ANALYSIS TO ANTICIPATE BITCOIN VALUE |

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

| | |
|---|---|
| Date of approval: | 15 Jun 2022 |
| Project Reference Number: | P137852 |

# Appendix C - Source Code