UNIVERSITY OF
WESTMINSTER
FACULTY OF
SCIENCE & TECHNOLOGY

The Analysis of Social Media Influence on Cryptocurrency Prices

by

LUKE SCHARTAU

(W1667774)

Supervised by

TASOS PTOHOS

Submitted in partial fulfilment of the requirements of
the Dept. of Computer Science
of the University of Westminster
for award of the Master of Science

MAY 2018

DECLARATION

I, Luke Schartau declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This has not previously been accepted as part of any other degree submission.

Signed : L. Schartau

Date : 12/09/2018

FORM OF CONSENT

I, Luke Schartau hereby consent that this Project, submitted in partial fulfilment of the requirements for the award of the MSc degree, if successful, may be made available in paper or electronic format for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed : L. Schartau

Date : 12/09/2018

Abstract

This paper analyses the influence social media is having on Cryptocurrency prices across platforms such as Twitter and YouTube. The data collected across these platforms is regarding cryptocurrencies such as Bitcoin, Ethereum, Ripple and Litecoin. The overall data collection spanned over a four-day period in total. To determine the effect social media is having if any, both statistical and sentiment analysis were used to detect the number of tweets and videos including the cryptocurrency name. Whilst, analysing the language used determining whether there is a higher volume of either positive or negative polarity scored. The paper reviews issues such as accessibility of social media data and the accuracy of sentiment analysis with the use of reviewed literature. From analysing both the data collected, and literature studied, Bitcoin and Litecoin showed some correlation between the social media data and prices of both. This was due to prices increased and they both scored a higher positive polarity overall. In addition, Ethereum and Ripple displayed a weak correlation between the overall price and polarity. This was due to prices decreasing and their polarity scores detection higher positive language overall.

Acknowledgements

I would like to thank my project supervisor Tasos Ptohos, for the support and guidance throughout the project especially in my early stages of planning, critically analysing my project to help me develop it further.

Table of Contents

List of Tables

Table of Figures

1. Introduction

1.1 Problem Domain

In recent years, there has been an increasing hype towards using Cryptocurrencies and the use of blockchain technology. A vase amount of this started from the Cryptocurrency Bitcoin (Satoshi, Nakamoto, 2009), which has fluctuated in prices dramatically (Jermain Kaminski, 2014) along with new Cryptocurrencies emerging from this revolution. A lot of information about these Cryptocurrencies is being shared on social media platforms and individuals are keen to share their ideas on the matter. With the market prices of these currencies changing on a regular basis this project aims to look at the influence in which social media has on the Cryptocurrency Market Share Prices.

Previous work shows (Phillips, RP., Gorse, DG., 2017) an interest mainly around the effect of Twitter on the Bitcoin market share price (Kim et al., 2016). This project will build on previous research by aiming to look at a selection of Cryptocurrencies, whilst using an array of social media platforms in order to gain a more in-depth analysis (Matta, Lunesu, Marchesi, n.d). This will provide details to what level of effect social media is having on the Cryptocurrency market.

Some of the main social media platforms being used today are Twitter, Facebook, LinkedIn and YouTube. These should be referred to as a social media ecosystem (Hanna, R., Rohm, A., Crittenden, V., 2011) and “stop treating them as individual elements” (Hanna, R., Rohm, A., Crittenden, V., 2011) as they share many important factors. Recent work by Evangelos, K., Efthimios, T and Konstantinos, T., (2013) and Sitaram Asur and Bernardo Huberman, (2010), shows the large power in which these social platforms hold for predicting future markets. They show flaws into how organisations are not carrying out the right analysis on the data, putting forth the idea of utilising sentiment analysis for this process.

Whilst this poses all good, Sabrina Bresciani and Andreas Schmeil, (n.d) indicate the new issues in which these social media platforms are facing, two in particular is “monitoring/ managing the truthfulness of information” and “taking into account cultural differences and preferences”. This shows there is still needs to be a balance in the use of social media data due to issues regarding data quality and accuracy. To expand on existing research, the project will focus on four of the top ten (CoinMarketCap, 2018) cryptocurrencies currently on the market which are Bitcoin (Satoshi, Nakamoto, 2009), Ethereum (Ethereum Project. 2018), Ripple (Ripple. 2018) and Litecoin (Litecoin. 2018).

1.2 Motivations

Motivations towards to this research area are due to Cryptocurrencies are a fairly new market and currently research is mainly towards predicting the future prices around Bitcoin. So, this project will focus on more than one cryptocurrency and the effect of social media data which will broaden research towards this area. Additionally, this project will expand the author’s skill

set and learn more about Cryptocurrencies and the huge possibilities which social media data presents.

1.3 Aims

The aim of this project is to investigate into whether social media is having any potential influence on Cryptocurrency prices. The project aims to analyse data over a selected period of time from social media platforms (Twitter, Facebook, YouTube and LinkedIn) looking at variables such as number of mentions and vocabulary used for some of the most popular Cryptocurrencies available (Bitcoin, Ripple, Ethereum and Litecoin). The analysis will provide feedback into any correlation between the two datasets and show in specific moments whether popularity of discussion for these Cryptocurrencies had a knock-on effect in their prices.

1.4 Objectives

When constructing the objectives for the project, Bloom's Taxonomy of Learning Domains theory (Bloom's Taxonomy, 1956) has been considered along with remembering to apply the SMART (MindTools, n.d.) metric to the objectives. The objectives have been set out in order to achieve the project aim, they are as followed:

1. Identify different approaches of gathering data sources.
2. Select variables from each social media platform to be used as a measurement (keywords, number of mentions, etc).
3. Process the data using Apache Spark ready to analyse.
4. Apply sentiment analysis to determine whether a Cryptocurrency has more positive or negative discussion via social media.
5. Analyse the data using statistical analysis to compare which Cryptocurrency has the most popularity of discussion via social media.
6. Discuss any issues with the quality of the data (e.g. contextual understanding).
7. Evaluate if there is any correlation between social media's discussion of the selected Cryptocurrencies and their prices.

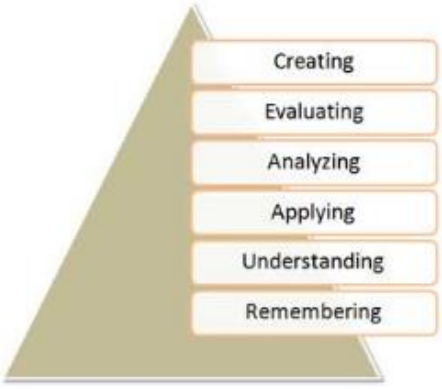


Figure 1 - Bloom's Taxonomy of Learning Domains (Bloom's Taxonomy, 1999)

1.5 Methodology Summary

The project methodology and approach will be detailed in (Methodology) chapter. A brief outline of the structure can be found in (Appendix B – Project Methodology). The idea being to collect and store social media via R studio (RStudio, 2018), then formatting and processing the data ready for analysis. The two analysis methods chosen are statistical and sentiment analysis. Statistical analysis (SAS UK. 2018) will be used to evaluate the number of tweets and mentions within the social media platforms. Additionally, sentiment analysis (Brandwatch, 2015) which measures the emotion and opinion used about a specific topic (Brandwatch, 2015). This will be used to detect various emotions and as well positive or negative comments.

1.6 Project Success

From the completion of the project, it must be determined whether the project has been a success. In order to do so all objectives must have been met. The first objective will be measured whether a selection of data sources has been presented, and a clear choice of data source has been selected, stating the reasons why. The second objective will be met when a table has been created determining the variables which are going to be collected and measured. The third will be met when the data has clearly been processed using Apache Spark. To determine whether objectives, four and five has been met visualisations will be created to show this analysis easier. Objective six will be met once a detailed analysis has been reported regarding the data quality. Finally, objective seven will be met by placing two graphs one from each dataset and each Cryptocurrency to say whether social media has an effect on the prices and if that is a positive or negative effect.

1.7 Risks and Back-up Resources

Some of the risks which the project might encounter would be the availability of data and the accuracy of data sources. The first data source is the Cryptocurrency prices downloaded from Kaggle.com, this dataset is scraped from an API by someone else. Data accuracy will have to be checked across the dataset in order to check there are not mistakes. A back-up resource for the Cryptocurrency prices to independently access data through either CoinAPI.io (CoinAPI, 2018) or CoinMarketCap API (CoinMarketCap, 2018). A similar risk might occur when using the social media platforms API's to extract data. Data accuracy needs to be remembered and checked, due to some social media platforms offer a limited extraction of data to developers with their free API's instead of Enterprise level. If any social media platform was to restrict their API usage, a method of finding a pre-extracted dataset would be opted for.

All the software programs being used within the project are open-sourced so there is not pre-set time limit on the amount of usage allowed. As the project requires some new programs to be learnt, this is another factor which needs to be considered as some programs may take longer than expected. To avoid this risk as much as possible, tools have been selected which

have some very basic knowledge and offer a large variety of learning resources either through the university or online. A back-up is in place to carry out the sentiment analysis on the data as this is a crucial stage in the project. If sufficient time if left at the end of the project this tool will be used to compare sentiment analysis results from the data. The back-up analysis plan is to use Brandwatch Analytics (Brandwatch, 2015) which offer a three-month access to their software which can perform sentiment analysis through social media platforms.

1.8 Professional, Social, Economic and Legal Issues

Throughout this project various laws and standards will be kept in order for the project to be successful. Professional guidelines which will be followed will be from the British Computer Society Ltd (British Computer Society, 1957) which is their BCS Code of Conduct (British Computer Society, 1957). Economic issues are very low on this project as many software's being used are outsourced and open source. Laws and standards which need to be recognised are *Data Protection Act 1998* which covers various standards in dealing with personal data. Another standard which will come into place on May 25th, 2018 is the *General Data Protection Regulation (GDPR).* Any data will be anonymised, and data will not be kept for any period longer than needed. Another couple of laws and standards which will be obeyed are *Computer Misuse Act 1990, Intellectual Property Act 2014* and *Copyright Act 1956.*

1.9 Project Management

The project will follow the task schedule shown as a Gantt chart. The task schedule outlines the major tasks needed to be completed within the project. Sufficient time has been made for learning new skills and for the analysis stages as these need to be prioritised. The task schedule will be used a reference throughout the project to determine if the project is behind or in front of the proposed project plan. The reason for producing a simple gantt chart is due to there is only one individual completing the project, so they know what exactly needs to be done and are involved in every aspect of the project. If more individuals were involved in the project, then a detailed plan would need to be completed as it would be passed on to different areas within a business or project. The Gantt chart can be found in (Appendix A – Project Schedule).

1.10 Project Road Map

The chapters following the introduction of the report start with the Literature Review. The literature review covers topics of interest relating to the projects focus and help to give a broader understanding of cryptocurrencies and social media and in addition the problems these may face. Continuing, the next chapter is the Methodology which outlines the initial plan for the project and the chosen methodology with a discussion justifying reasons for the change of direction. It also provides a detailed report of the variables being collected and the packages used in order to succeed with the individual tasks. Data Formatting and Cleansing follows which

goes through the steps taken in order to cleanse mainly the textual data collected. After this, the Analysis chapter follows which includes a thorough report of the results found for each cryptocurrency across the social media platforms. Issues and challenges faced in project follows this chapter, discussing general problems faced throughout and data quality issues. Finally, the Evaluation chapter is last in the report with outlines the conclusion to the results which compares with others in the same research field. In addition, states the limitations and potently future work for the project determining whether the project was a success or not.

2. Literature Review

2.1 Introduction

This chapter of the report discusses the literature which has been researched to help give a more in-depth knowledge about various topics relating to the project domain. These topics are as followed, first an introduction into cryptocurrencies followed by social media platforms being used. In addition, the advantages and disadvantages of social media will be discussed with the accessibility of social media. Then, the accuracy of sentiment analysis followed by the future of social media, cryptocurrencies and predicting future markets.

2.2 Introduction to Cryptocurrencies

Cryptocurrencies have increasing made huge developments not only in the technology they use, but in the ability to still be growing day by day as an industry. This year in 2018, Yuan and Wang (2018) reported that there are now around one thousand five hundred cryptocurrencies currently in the market with the total market cap to be around five hundred billion dollars. Saad and Mohaisen (2018) explain why this might be, by stating that every new cryptocurrency is improving on the previous faults of another coin not only improving scalability but programming ability. It would be fair to say not many people believed the first cryptocurrency Bitcoin, (Satoshi, Nakamoto, 2009) would still be around nearly ten years on. Though Bitcoin (Satoshi, Nakamoto, 2009) was not the first digital currency, Bohr and Bashir (2014) explain in their research how electronic currencies such as “DigiCash and CyberCash were innovative in their ability to digitally transfer large amounts of money” (Bohr and Bashir, 2014) but these were not implemented and only offered some privacy.

He, et al. (2017) would agree that cryptocurrencies are used as an investment similar to stock exchanges and with more and more places around the world accepting them as valid payments it does not seem like cryptocurrencies are disappearing anytime soon. Though it has not been a smooth ride for cryptocurrencies, Bitcoin especially acting as the test pilot within the industry being linked with the dark web such as “Silk Road”. The argument was Bitcoin was associated with “facilitating the exchange of illegal goods around the world” (Bohr and Bashir, 2014) and this being the problem by not being able to easily track cryptocurrency transactions. Cryptocurrencies have many similar problems, as Bohr and Bashir (2014) would suggest that regulators have the extended job of discussing whether current cryptocurrencies fall under anti-money laundering laws now. The outcomes to these discussions can have a huge impact on the prices and would leave uncertainty within the market until figured out. Nguygen (2016) proposes the argument against regulators saying that their tight rules and regulations do not make the financial sector any safer than before. The problem the world faces is because cryptocurrencies are decentralised, governments imposing regulations on them takes away some control from the communities.

Most if not all cryptocurrencies use a technology called Blockchain also known as a distributed public ledger. Blockchain is “maintained by a decentralized computer network, where all online transactions are recorded, and everyone is allowed to connect, to send or verify transactions” (Nguyen, 2016). This technology allows the ability for virtually anonymous transactions. Mukhopadhyay, et al. (2016) explains that transactions start when a buyer sends some cryptocurrency to the second user, a process called mining takes places and verifies the transaction taking place which adds to the blockchain. This is supported by Vujičić, Jagodić and Ranđić (2018) which reveal that to allow more coins into circulation in the blockchain, “the first transaction in each block creates a new coin” and every transaction for security reasons has its own unique hash. Blockchain itself is moving forward itself, with Blockchain 3.0 being released improving on existing versions. Blockchain 3.0 is not just focusing on its main use for cryptocurrencies as Saad and Mohaisen (2018) point out the technology has applications in various areas such as “distributed censorship resistant organization models, digital identity verification and decentralized domain name system” (Saad and Mohaisen, 2018).

With every new technology there can be some underlying issues still to be tackled and new methods of online criminality taking place. One way a threat to the blockchain might happen is through an attacker creating “multiple nodes and try to validate an invalid transaction” (Mukhopadhyay, et al., 2016) but blockchain tackles this making it essential that all miners carry out what are known as “resource intensive task”. He et.al (2017) addresses the issue of the cryptocurrency communities using private keys, if these keys are not looked after or managed cryptocurrencies can be lost or stolen from others, stating that at the end of November 2017 that Bitcoin had announced around thirty-one million dollars had been stolen.

2.3 Social Media Platforms

In today’s society, social media has become part of billions of peoples’ everyday lives. Facebook (Facebook for Developers. 2018) and Twitter (Twitter Developers. 2018) being described by Wang, et.al. (2016) as the “go-to platforms” for users wanting to be involved in the social media era. Shahnawaz and Astya (2017) reveal how social media platforms are being used for various activities such as sharing personal life events to one another. Whilst, having the ability to engage in conversations with a number of friends and express their likes and dislikes about services and products. Social media platforms provide an array of activities and features for users to get involved with their sites, one example is providing the ongoing trends, which show the most talked about topics currently on the site, which Trupthi, Pabboju & Narasimha, (2017) express in their research labelling them “microblogging sites”. Joshi and Tekchandani, (2016) define this as social media/ blogging sites which cap the number of words a user is allowed to express their opinions globally through the sites.

Additionally, YouTube (Google Developers. 2018) is leading the social media platforms in providing a global interface for sharing video content. Bhuiyan, et al. (2017) outlines YouTube’s capabilities by explaining the platform allows users to add comments and ratings to videos as

informal feedback for the creators. Users can upload videos on near enough any topic they choose within reason. With all these social media platforms, vase amounts of data is produced every second from activities on the sites, this data arrives in various formats, also known as Big Social Data. Ibrahim, et al. (2017) voices how this big data can be used to gain valuable insight about businesses and organisations and many other topics. Both Jain and Katkar (2015) and Wang, et al. (2016) agree that this enormous variety of data can be an exciting time for researchers and analysts discovering new in-depth analysis from this, one of which describes it as a “gold mine” for those who can see the potential.

2.4 Advantages and Disadvantages of Social Media Platforms

Social media platforms can provide a great source of data for analysis in many areas as discussed earlier in this paper. Chatterjee and Goyal (2015) stress how Twitter (Twitter Developers. 2018) especially is an excellent source for gathering various opinions about most topics. Whilst also, emphasising how these social medias platforms are so diverse from governments to celebrities, growing day by day. On the other hand, Vela, Martinez and Reyes (2012) argue that about thirty percent of the population have access to the internet. Meaning that social media data cannot always truly understand how the whole population feels about a specific topic.

Additionally, due to social media platforms allowing users to express opinions daily with no costs also reinforces why Sir Tim Bernard’s Lee (Tim Berners-Lee. 2018) created the internet to express ideas freely. Platforms can be seen to improving freedom of speech providing advancement on human rights, which is laid out in research by Vela, Martinez and Reyes (2012). With expressing opinions on the web and allowing anyone to view this content has its own problems. Deodhar, Divakaran and Gurusamy (2017) state how users information could be stolen by criminals and could influence fraud if not addressed properly. Mahmood (2012), builds upon this statement in talking about how attackers could steal identities linking them to false email addresses and bank statements. These risks can be categorised into three main topics, Joseph (2012) describes these as “communication, operational and political” with the most popular flaws being communicational linking to overall management of social media accounts.

Furthermore, Misra and Such (2016) present the importance of privacy controls and how social media platforms are facing the underlying challenges in giving users sufficient controls with their personal data whilst tackling the issue that users have different levels of friendships with different individuals using the platforms. As well as, this social media can open online communications between individuals but close off physical communications due to users being on their electronic devices too much as Vela, Martinez and Reyes (2012) suggest.

2.5 Accessibility and ethics of Social Media Data

As shown above, there can be many advantages and disadvantages in using social media platforms. Many users of these platforms can sometimes forget the consequences in which social media platforms can provide in terms of personal data being publicly available to anyone using them. As stated, many researches and data enthusiasts find social media data to be very valuable and allows public opinion insights. Deng (2017) would agree that ethics in publicly available social media data need to be highly considered and thought about, likewise Patti, Damiano and Bosco (2017) expand on this by stating that social media data can cross paths both publicly and privately and can be difficult to distinguish between the two. A number of companies collect social media data such as user activities such as the topics of discussions and their personal interest. Deodhar, Divakaran and Gurusamy (2017) draw upon this enouncing that due to even though these communications are public they could still be classed as “privacy breaches”. Arguably, users could avoid posting information they do not want the general public to see. Small et al. (2012) proposes users delete certain pieces of information they might give away too much such as locations. Though as explained by Small et al. (2012) users still run the risk of their information being archived, which would could problems all together.

At the same time, Morstatter, Liu and Zeng (2012) would agree that sharing public data can vastly improve specific fields of study and inspire the new generation of researches to get involved with analysing this public data. Plus, pointing out the social science communities use this data to help push forward academic papers and avoid repetition. On top of this Morstatter, Liu and Zeng (2012) research experiment shows how these social media platforms are starting to restrict access for third parties and restricting what sort of public data can be collected and the quantity in which individuals can collect. Small et al. (2012) tackled similar problems when accessing public data via Twitter (Twitter Developers. 2018), where Twitter enforced their access to be cut due to requesting too many calls. This can cause public social media datasets to have inconsistency in their data and potentially jeopardise results. Dealing with this issue can be difficult as Morstatter, Liu and Zeng (2012) proved as they presented the idea every new reader of a research paper involving social media data should carry out a new data collection using the same methods. This method in itself can prove very unreliable due to the data could be completely unrelated effecting the results.

2.6 Accuracy of Sentiment Analysis

Sentiment analysis is what is described as determining the emotion or opinion used by a user about a specific topic. Sentiment analysis is broken down into two categories polarity, which determines whether the data is positive, negative or neutral or the second category sentiment, meaning the emotions used (Shahnawaz and Astya, 2017). This can also be described as “different levels of granularity” by Patti, Damiano and Bosco (2017) which later outline how the basic sentiment is to determine the positive and negative scores. Though initially applying

sentiment analysis to textual data can be quite difficult and non-accurate due to the language used such as slang words and misspellings can cause issues as expressed by Deng (2017). Bhuiyan, et al. (2017) asserts how importance pre-processing can be to remove any stop words and irrelevant language used. Kaur, Mangat and Nidhi (2017) indicate that sentiment analysis requires subjective text in order to determine a positive or negative score, so removing text which states facts is essential.

For sentiment analysis there are three main methods in order to determine the score of words. This can be shown in the figure below. Abirami and Gayathri (2016) explain how dictionary-based methods are both the most common and simplest way to perform sentiment analysis, whereas Kaur, Mangat and Nidhi (2017) make note that this method can struggle in discovering context or "domain-oriented opinion words".

	Approach	Advantage	Limitation
Machine Learning	Make use of supervised and unsupervised techniques to perform sentiment analysis.	It can be customized to deal with different domains.	Classification is only limited to 2 categories such as "positive", "Negative".
Dictionary Based	It make use of reference dictionary to classify the individual sentence.	Computation overhead is lesser as there is no training of dataset.	It is not suitable for context /domain specific classification.
Ontology Based	Used in feature extraction stage of sentiment analysis.	Mainly take cares of semantic relationship between the features.	Updating ontology is difficult task

Table 1 - Abirami and Gayathri (2016)

Priyanka and Senthikumar (2016) explains one of the most commonly used dictionaries is "SentiWordNet" which is a database downloaded online and can be applied to any textual data of choice. Wang et al. (2016) would agree that the "lexical-based" (dictionary) methods provide a more in-depth analysis as they provide a range of emotional scores unlike machine learning methods whilst costly considerably less. Woldenmariam (2016) claims that results show the lexicon-based method being down on accuracy by 9.88% through polarity scoring against "Recursive Neural Tensor Network model", though showing a higher accuracy in positive recognition. Perikos and Haatzilygeroudis (2018) additionally state how measuring the accuracy of sentiment analysis methods can be very time consuming and by near enough impossible on large datasets manually.

Furthermore, sentiment analysis faces deeper challenges in understanding what a user is intending to say in their statement, as discussed by Trupthi, Pabboju and Narasimha (2017) users can express their thoughts in various ways which adds complexity to the problem. One of the biggest issues with sentiment analysis is dealing with contextual understanding such as sarcasm. For example, Patti, Damiano and Bosco (2017) back this up by talking about how sarcasm can cause false understanding to the system, providing inaccurate results calling it

“polarity reversal phenomenon”. This too can be supported in Abirami and Gayathri (2016) research which explains how this “polarity shift” can be difficult to deal with. The word “don’t” in a sentence is the polarity shifter, picking this up can boost accuracy.

Another problem with the accuracy is dealing with classification. Classifying issues arise when systems or libraries take the opinionated data and falsely categorise the word as discussed by Shahnawaz and Astya (2017). Exceeding this, Abirami and Gayathri 2016 see the issue in classification to be extended in the limitation of only having either positive or negative categories for textual data. Proposing the answer to improving accuracy should stretch as far as five categories such “strong positive and positive”. Liang and Dai (2013) would argue this, showcasing their experiment results as receiving 90.17% by only using two categories of polarity to increase accuracy. This in itself can actually have the opposite effect as data is categorised too generally.

Again, classification problems are seen as a major factor determining accuracy, Jose and Chooralil (2016) stress how classifying each words sentiment can cause positive results mainly to be inaccurate. Another problem sentiment analysis faces as Save and Shekokar (2017) points out is that a word can have two different meanings, understanding the right context can be difficult. So, it is vital if using machine learning techniques training the data sufficiently is essential. Continuing on from this, many see the issue not only in the language used by the source of textual information and its limitations on users’ opinions. Abirami and Gayathri (2016) label this as “data sparsity” detailing how social media platforms restricting the number of words allowed for a user to express themselves, can impact the selection of language used making it more difficult to analyse this data. To improve accuracy of sentiment overall extending the amount of data being analysed can improve accuracy (Priyanka and Senthikumar, 2016).

2.7 The Future

Predicting Future Markets

As discussed, social media data with sentiment analysis can provide interesting and actionable insights into the general publics’ emotions and opinions across various topics. A popular area of research is predicting future markets such as stocks or cryptocurrencies using social data. Unlike Radityo, Munajat and Budi (2017) who claim more traditional methods of predicting future markets is through technical analysis, using a method called “NEAT” which analyses a specific time in history and applies to future dates. Twitter is one of the most common social platforms used in predicting markets due to its variety of discussions and the number of users it has. Phillips and Gorse (2017) argue that Reddit is potentially the best social platform for predicting markets especially cryptocurrencies, revealing that it focuses more on in-depth conversations between users about various ideas and not just generally comments. There has proven to be some correlation between social media activities in predicting markets such as cryptocurrencies as supported by Lyudmyla, Vitalii and Tamara (2017).

Future of Cryptocurrencies

The future of cryptocurrencies still cannot be full guaranteed in how far they will stay around due to various variables affecting this result, regulations being a big part of the issue. In contrast, Bohr and Bashir (2014) estimate the future of virtual currencies to be strong as they have large communities supporting them through their struggles in the market. Though, it can be stated that the technology blockchain has a more promising future, as discussed by Nguyen (2016) banks are investing into how they system can use this technology to remove as much competition as possible from cryptocurrencies. Another factor determining some cryptocurrencies futures specifically Bitcoin's (Satoshi, Nakamoto, 2009), as Radityo, Munajat and Budi (2017) remind is its limited supply of coins and this could also have impact of the prices of coins are this moment comes closer.

Future of Social Media

In 2011, Baatarjav and Dantu gave a summary on the future of social media platforms stating the focus would need to be addressing privacy issues and explaining the real changes will be the ways users share content and to who exactly they choose to share their content with. This can be agreed with seven years on, that privacy issues have become the most important topic of social media platforms, with platforms constantly improving the accessibility of privacy controls to users to make them feel more at control with their online social interactions. This will be an ongoing issue for the future social media platforms.

2.8 Summary

From the literature reviewed within this chapter, the main problem social media data presents are first of all its accuracy regarding the method of analysis. Additionally, the accessibility of the data seems to be an issue on the rise due to social media platforms increasing their restrictions to data and improving privacy policies for users.

3. Methodology

3.1 Introduction

This chapter provides a comprehensive report on the approach taken for this project. The methodology outlines each step being taken to ensure the successful of the project. Details include of the type of data being collected, storage specifications with the packages being used to provide formatting and sentiment analysis. A discussion about the changes made to the initial approach and specified at the end of the chapter.

3.2 Gathering Data

The first stage in the methodology is to gather both the social media data and cryptocurrency data. As mentioned in the objectives (1), the plan is to identify various ways of collecting the data and identify the best methods. The initial plan is to carry out batch processing to enable both data sources to be downloaded and not live streamed/ processed. The reason for this choice will allow a selected period to be analysed and batch processing will still be a new skill to learn for the project to be a success.

The initial plan for gathering the Cryptocurrency data was to use an existing dataset from Kaggle (Kaggle, 2018), this data has already been scraped from CoinMarketCap API free to use. It covers historic information on open, high, low and close values for all cryptocurrencies including Bitcoin, Ripple, Ethereum and Litecoin. Other variables such as close ratio, min and max values for each day for each cryptocurrency. The dataset can be found from the following link (*https://www.kaggle.com/jessevent/all-crypto-currencies/data*). When this data was inspected more, it only displayed the closing price for example at the end of each day not every hour. It was decided to improve analysis and to provide more opportunity in seeing a connection between tweets and prices, a dataset with closing prices of every hour would suite best. The methods of using packages in R studio (RStudio, 2018) called “crypto” (J. Vent, 2018) and “coinmarketcapr” (AbdulMajedRaja, RS and Srivathsan K, 2017) which use API’s to crawl data was outweighed again. The final decision to use datasets from “http://www.cryptodatadownload.com” was made. This decision was made due to CryptoDataDownload (CryptoDataDownload, 2018) providing the closing price and more variables of every hour for all four cryptocurrencies being researched. The datasets provide prices from each currency into United States Dollars from the “Kraken.com” exchange. The choice for this exchange is due to it providing prices for all coins and to improve accuracy of results, so data is not sourced from several exchanges and currencies.

The initial plan for gathering the social media data would be provided from platforms such as Twitter, Facebook, YouTube, LinkedIn which would be scraped from API’s to gather historic data through R studio (RStudio, 2018). Each of the social medias provide API access, though from the initial plans Facebook restricted their API access and required an additional data management and video to be submitted in which Facebook decide whether to approve or not.

Facebook (Facebook for Developers. 2018) additionally, closed the “search” feature on their API (Facebook for Developers. 2018) which was the main source of data scraping from the platform. Furthermore, after carefully consideration LinkedIn (LinkedIn. 2018) was chosen not to be analysed as its API access it limited due to the focus of it being on one individual’s data. The API does not provide a search feature for posts only search Companies, Jobs and People as the “Rlinkedin” (Michael Piccirilli, 2016) package details. Other social media platforms were researched such as Instagram and Reddit, these both did not provide enough data to scrape via their API’s.

As the initial plan changed, using the R package “SocialMediaLab” (T. Graham., R. Ackland, 2017) which provides “a suite of tools for collecting data from social media” (T. Graham., R. Ackland, 2017). As well as, package “Rlinkedin” (Michael Piccirilli, 2016) which helps connect to the LinkedIn API will be used in this process would not be necessary. For accessing Twitter (Twitter Developers. 2018) data, the “standard search API” would be used in line with the R package “rtweet” (0.6.7) (M. Kearney, 2018) to search for the cryptocurrency name within tweets. One main feature is the package allows up to 18,000 tweets to be collected every 15 minutes as Twitter restrict the number of calls. The twitter data collection will not include any retweets as it would just be repeating results and this research focusing on the original tweets. For collecting YouTube data, the package “tuber” (0.9.7) (G. Sood, 2018) would be used as this provided a similar feature of searching for published videos containing keywords of choice.

3.3 Storing Data

Once, the data has been gathered it will then need to be instantly stored. The Cryptocurrency dataset is in CSV format and between 649KB and 692KB dependent on the selected cryptocurrency. This will not require any additional storage as this will be downloaded straight on to a laptop. The method of storing the social media data is in the workspace R studio (RStudio, 2018) provide. R will be used to convert the data into a data frame or matrix to be ready for formatting and cleansing. After this, the data will either be converted to CSV format using package “jsonlite” (Ying Chen, 2016) or using the package “rtweet” (0.6.7) (M. Kearney, 2018). It will be easier to initial process and cleanse. If this method proves too difficult or the data frame size is too large the JSON files will be stored using MongoDB (MongoDB, 2018).

3.4 Cleansing Data

Cleansing the data can be done in two ways. The Cryptocurrency dataset will be cleansed in R, this will remove any information which is not about the four currencies selected. Once the selected period has been chosen, the dataset can then remove any unwanted dates from the four currencies and can be saved under a new csv file. The initial plan to cleanse the social media data was to cleanse the data within Apache Spark (Apache Spark, 2018) using Hortonworks (Hortonworks, 2018). As Spark is not being used, the data formatting and cleansing will take place in R studio (RStudio, 2018). The package “lubridate” (V. Spinu et al.,

2018) version 1.7.4 will be used to deal with date formats and changing where appropriate specifically for the csv dates. Another package called “qdap” (B. Goodrich, D. Kurkiewicz and T. Rinker, 2018) version 2.3.0 will be used to remove language not appropriate for analysing. Variables such as removing hashtags, URLs, twitter mentions and stop words can be performed using this package.

3.5 Processing the Data

The original idea for processing the data was to download Hortonworks Sandbox for HDP (Hortonworks, 2018) and use Apache Spark (Apache Spark, 2018). After consideration, it was decided to use R to process the data. The reason for this choice is due to previous research revealed that R can handle processing the social media data well and have proven to be successful in previous projects without complicating the project too much. This method will reduce problems occurring due to the data will be collected, stored, cleansed, processed and analysed all in R. The cryptocurrency data will focus on the closing price of each cryptocurrency and this is the most important variable given in the dataset. This will allow the ability to see which currency has risen in price or which one has dropped the most over the selected period. It will also process which cryptocurrencies has the highest number of mentions, the language used and its sentiment.

3.6 Sentiment & Statistical Analysis

Once the data has been processed it will then be analysed to discover any patterns and correlations between the two data sources. The Cryptocurrency prices dataset will be analysed to see the rise and falls in the four cryptocurrency prices over time. This will be plotted on various graphs especially line graphs to easily show the trends in prices. The Cryptocurrency social media data will be analysed using statistical analysis first looking at the number of mentions across both Twitter and YouTube in the selected time. This data will be plotted onto a graph and presented side by side with the Cryptocurrency prices to see if any correlation is clear.

The next variable to measure was to measure the sentiment score of the language used for all four cryptocurrencies via Twitter and YouTube. The initial plan was to use a list of keywords to determine a selection of emotions for each currency, three to five words would be selected for each emotion, and around five to six emotions. After consideration and to improve accuracy of the data it was decided to just search for the cryptocurrency name and then analysis the whole sentence it was used within to determine its sentiment score. The package “syuzhet” (1.0.4) (M. Jockers, 2017), would be used to detect polarity scores of the language used with the keyword, which determines either positive or negative.

Additionally, the package can calculate sentiment score of all words using the keyword categorising them into eight different emotions “anger, anticipation, sadness, trust, joy, disgust,

surprise, fear". "Syuzhet" (M. Jockers, 2017), works as a lexicon-based dictionary and is developed in a "Nebraska Literary Lab" (M. Jockers, 2017). This method will provide more accurate results than searching five words relating to one emotion and repeating these tasks. Though as discussed in (Accuracy of Sentiment Analysis), sentiment analysis can be inaccurate at times, but this method is best suited to the time frame and resources available. The analysis will investigate what emotion is scored highest within the language used and the polarity score of the social media language. Again, to compare this analysis will the prices dataset, a graph will be put side by side and another graph will integrate both values to see if there is any correlation between prices and emotions.

Continuing, two other variables initial was planned to be measured, after further research it was discovered that measuring emojis was not allowed with the Twitter standard API. Likewise, the location of tweets was going to be analysed, it was revealed that a vase number of users to not input their location via Twitter and this data would prove to be inaccurate and display too many nulls.

3.7 Visualisation and Evaluation

The last stage in the project cycle is the evaluation of results from the analysis to determine whether social media has any effect on Cryptocurrency prices. To make the evaluation easier, various visualisations will be produced to show the correlation if any between the two datasets. This will clearly show the results and determine the outcome. As mentioned, this would be placing two graphs next to each other as well as integrating graphs. Another type of visualisation used for keywords relating to each Cryptocurrency will be term frequencies and term co-occurrences. Both R-Studio (RStudio, 2018) and Tableau (Tableau, 2016) will be used to visualise the results. From the analysis, both good and bad trends will share a similarity in how easy they are to predict from the results, due to individuals use social media platforms to express both their good and bad thoughts about topics, in this case Cryptocurrencies.

3.8 Discussion of Changes Made

The methodology detailed above outlines specific changes which have been briefly explain to why these changes were made. This section will justify these alterations to the project. One of the changes made was the source of the cryptocurrency price dataset, as stated this change was made to allow analysing by the hour instead of day. This change should hopefully make it easier to see if any correlation if happening between the two datasets. As discussed, Facebook (Facebook for Developers. 2018) and LinkedIn (LinkedIn. 2018) were not analysed due to API access been restricted in the features available and existing features not suiting the project well.

One of the big changes to the project methodology was the removable of Apache Spark (Apache Spark, 2018). This was removed as more research on existing and similar projects took place, it was discovered that R had previously shown enough capability to analyse and

process the social media collected. Removing this also allowed more time to be focused on learning R and experimenting more with visualisations. Finally, the last change to be made was the way in which sentiment analysis was performed. The chosen method displays more accuracy than the initial as all tweets and video titles would be calculated than included the cryptocurrency keyword.

3.9 Summary

From the chosen methodology chapter details about the different ideas of collecting data sources are specified. Along with details about the variables and measures being collected via R (RStudio, 2018). Details about the method of statistical and sentiment analysis are stated.

4. Data Formatting and Cleansing

4.1 Introduction

In this brief chapter, details will be given for the steps taken in order to format and cleanse the data ready for analysis. Print screens are shown to show the effect of each step taken using a package in R (RStudio, 2018).

4.2 Data Collection

The first figure below shows an example of the data collection process using the Twitter API (Twitter Developers. 2018). As shown, the package requests the data from the API when the maximum is reached, it requests again until the total amount has been collected, for this example 300,000.

```
> Litecoin1 = search_tweets(q = "litecoin", n = 300000, type = "recent", lang = "en", include_rts = FALSE, retryonratelimit = TRUE, verbose = TRUE)
Searching for tweets...
This may take a few seconds...
Finished collecting tweets!
retry on rate limit...
waiting about 10 minutes...
Searching for tweets...
This may take a few seconds...
Finished collecting tweets!
retry on rate limit...
waiting about 11 minutes...
Searching for tweets...
This may take a few seconds...
Finished collecting tweets!
retry on rate limit...
waiting about 12 minutes...
Searching for tweets...
This may take a few seconds...
Finished collecting tweets!
Searching for tweets...
This may take a few seconds...
Finished collecting tweets!
```

Figure 2 - Data Collection via Twitter

The figure below shows an example of what the data looks like when collected from Twitter. The package “rtweet” (0.6.7) (M. Kearney, 2018), collects 88 different variables for every tweet. The two this project focused on was the “created_at” and “text” columns.

	user_id	status_id	created_at	screen_name	text	source
1	955860396006690817	1026553184834150401	2018-08-06 19:38:46	D_xheva	Why are you into crypto? #btc #ltc #eth #ada #xvg #bitc...	Twitter for iPhone
2	1016645809507655681	1026552483022168064	2018-08-06 19:35:59	Peterhughas	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
3	825785525374103552	1026552480455307264	2018-08-06 19:35:58	Leo1G2	Litecoin Adoption & Use Explained In Under 10 Min...	Twitter Web Client
4	950440947334205440	1026552288976855040	2018-08-06 19:35:12	ZokaPokaX	@litecoin_bull Make it a #rum party, if you're persistent I...	Twitter for Android
5	14601336	1026552282979000321	2018-08-06 19:35:11	CopyTraderCo	To the new guys & girls, let's see how tough you re...	Crowdfire - Go Big
6	14601336	1026549752148545541	2018-08-06 19:25:07	CopyTraderCo	Warren Buffet: 1 - 200x Leverage <U+2705> <U+25B6>...	Crowdfire - Go Big
7	952035808063688709	1026552006226268161	2018-08-06 19:34:05	CryptoNamer	ContentmentCoin #cryptocurrency #Bitcoin #Ripple #Et...	CryptoNamer
8	92292609976000512	1026551227029409792	2018-08-06 19:30:59	bitxapp	Current #crypto prices: #Bitcoin at $6,944, down -1.2%. ...	Bitx App
9	975649624240345088	1026550333055455232	2018-08-06 19:27:26	nasty777coin	Online lottery platform #ZeeringLotto Digital Currency S...	Twitter Web Client
10	957193561186828289	1026549611874271233	2018-08-06 19:24:34	RBoucer	Litecoin $LTC is valued at: $73.73 HURRY! Binance is curre...	dfgh5er41t98e19485t
11	2520368154	1026549232755306496	2018-08-06 19:23:04	ElixiumCrypto	What would he be wishing now? <U+0001F602> 1 - 10...	Crowdfire - Go Big
12	2520368154	1026546736699912192	2018-08-06 19:13:09	ElixiumCrypto	Investing Tips from a Pro 1 - 200x Leverage <U+2705> <...	Crowdfire - Go Big
13	717648725405872128	1026548586669006854	2018-08-06 19:20:30	yvonnewelch187	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
14	717648001200553984	1026548365301927936	2018-08-06 19:19:37	victorstewart33	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
15	717648001200553984	1026544160512212992	2018-08-06 19:02:54	victorstewart33	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
16	2799558638	1026547806167752706	2018-08-06 19:17:23	TaxTwerk	litecoin cryptocurrency review - bitcoin product review 0...	Videojeet
17	717644111889575938	1026547661606776832	2018-08-06 19:16:49	ianpaterson5151	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
18	238994600	1026547282039078913	2018-08-06 19:15:19	eagletwitt3r	WHEN LAMBO? No one really knows but you can reserv...	IFTTT
19	977242768044150784	1026547263382872066	2018-08-06 19:15:14	CryptoSigStat	"According to the statement on their blog, the aim of la...	Buffer
20	827345583815933952	1026547016350949377	2018-08-06 19:14:15	Lite_money	@Hydeez411 It days invalid address while I tried to send...	Twitter for Android
21	744479689205485568	1026546962575712256	2018-08-06 19:14:02	Social_mediasos	F.ree bitcoins here :) easy to use and refer friend and ma...	Twitter Web Client
22	1157511169	1026546735827406848	2018-08-06 19:13:08	WHardingKY	JPMorgan CEO Jamie Dimon Returns to Bitcoin Bashing,...	Crowdfire - Go Big
23	26448153	1026546616352747520	2018-08-06 19:12:40	johnkim77	Passion is like fire. If your passionate about something, ...	Twitter for Android
24	3316078646	1026546613790027776	2018-08-06 19:12:39	satoshiuncle_	Researchers Reveal Network of 15K Crypto-Related Scam...	satoshiword
25	4925308819	1026546079242678272	2018-08-06 19:10:32	borisrandall111	Top Trusted Crypto Currency Platform <U+0001F4B9> <...	Twitter Web Client
26	1018219726898716672	1026546011647356928	2018-08-06 19:10:16	Cryptomuginves1	Having fun talking about @omise_go in today's video. L...	Twitter Web Client
27	2261019318	1026545653005004800	2018-08-06 19:08:50	trade_den	Working on a few new things for the discord, we recent...	Twitter for Android

Figure 3 - Example data scraped via Twitter in R

4.3 Pre-processing

During the pre-processing stage, data was ordered by date and date formats were altered so all showed the same structure preparing for analysis. Additionally, the textual data was removed of its hashtags, URLs, tags, twitter URLs and stop words. As mentioned (Cleansing DataCleansing Data), “qdap” (B. Goodrich, D. Kurkiewicz and T. Rinker, 2018) version 2.3.0 was used to perform these steps ready for sentiment analysis.

As shown below, the first step was to make a copy of the tweets data frame to avoid mistakes happening to the initial raw and save time having to reload the workspace if something does not go right at first. The second line of code shows the process of ordering the twitter time variable to order from ascending to descending. Lastly, the rows were selected which fall within the four-day period being analysed.

```
###MAKE A COPY OF DATAFRAME
LITECOIN = Litecoin1

LITECOIN = LITECOIN[order(as.POSIXct(LITECOIN$created_at,"%d/%m/%Y %h:%m:%s", tz = "GMT")),]

###LITECOIN #8305 19655 = 11351
LITECOIN = LITECOIN[8305:19655, ]
View(LITECOIN)
```

Figure 4 - Date Formatting via R

4.3.1 Cleaning Tweets

An example of twitter data is shown below without any cleansing or format. Once the line of code to remove the hashtags is performed, the data frame must be changed into a matrix to allow the next step to process. The following print screens shows example of what the data looks like after each process is performed.

```
> head(LITECOIN$text)
[1] "Bitcoin @ £5,410.48 | Ethereum @ £240.14 | Litecoin @ £48.03 | Buy it online with bank transfer at https://t.co/Ckp
CoazTdY"

[2] "#Get_Cred makes it easy to get into #cryptocurrency. Download to #learn, #invest, &amp;amp; have some fun.\n\nAndroid:
https://t.co/uTe1GDhIRG \niPhone: https://t.co/U0KB6ZPguR\n\n@Bitcoin @EOS_io @ethereum @litecoin @NeblioTeam @wax_io @b
itcoincash @zencashofficial @KomodoPlatform #HODL https://t.co/VCsfx6qXax"
[3] "#BTC #bitcoin $6479.13\n#ETH #ethereum $279.98\n#BCH #Bitcoin Cash $535.74\n#ltc #litecoin $56.53\n#XRP #ripple $0.
34\n#XMR #monero $94.5\n#DSH #dash $141.5\n#criptovalute \nhttps://t.co/2XT7wuUIYG"

[4] "Last litecoin price on bitstamp: $ 56.47"


[5] "Dear #Trump #Haters &amp;amp; #Lovers\n\n#realDonaldTrump swears never an #Affair with #StormyDaniels\n\n#Video caught
him Flipping #Stormy $10,000 #Money #Bird\n\n#Love #Bitcoin #Ethereum #LiteCoin #BTC #Xrp #Cryptocurrency #ISIS #PETCO #
FOX #CNN #IRSNEWS #DOJ #FTC #FED #ZOO #news https://t.co/OJjjGsQzIK"
[6] "LTC to USD price $56.35 https://t.co/rVzBs6fBj4 #litecoin #ltc #cryptocurrency"
```

Removal of hashtags

```
[1,] "Bitcoin @ £5,410.48 | Ethereum @ £240.14 | Litecoin @ £48.03 | Buy it online with bank transfer at https://t.co/Ck
pCoazTdY"

[2,] "makes it easy to get into . Download to , , &amp;amp; have some fun. Android: https://t.co/uTe1GDhIRG iPhone: https://
t.co/U0KB6ZPguR @Bitcoin @EOS_io @ethereum @litecoin @NeblioTeam @wax_io @bitcoincash @zencashofficial @KomodoPlatform h
ttps://t.co/VCsfx6qXax"
[3,] "$6479.13 $279.98 Cash $535.74 $56.53 $0.34 $94.5 $141.5 https://t.co/2XT7wuUIYG"


[4,] "Last litecoin price on bitstamp: $ 56.47"


[5,] "Dear &amp;amp; swears never an with caught him Flipping $10,000 https://t.co/OJjjGsQzIK"


[6,] "LTC to USD price $56.35 https://t.co/rVzBs6fBj4"
```

Removal of tags.

```
> head(LITECOIN_2)
     [,1]

[1,] "Bitcoin @ £5,410.48 | Ethereum @ £240.14 | Litecoin @ £48.03 | Buy it online with bank transfer at https://t.co/Ck
pCoazTdY"
[2,] "makes it easy to get into . Download to , , &amp; have some fun. Android: https://t.co/uTe1GDhIRG iPhone: https://
t.co/UOKB6ZPguR @Bitcoin @EOS_io @NeblioTeam @KomodoPlatform https://t.co/VCsfx6qXax"
[3,] "$6479.13 $279.98 Cash $535.74 $56.53 $0.34 $94.5 $141.5 https://t.co/2XT7wuUIYG"

[4,] "Last litecoin price on bitstamp: $ 56.47"

[5,] "Dear &amp; swears never an with caught him Flipping $10,000 https://t.co/OJjjGsQzIK"

[6,] "LTC to USD price $56.35 https://t.co/rVzBs6fBj4"
```

Removal of URLs.

```
> head(LITECOIN_2)
     [,1]

[1,] "Bitcoin @ £5,410.48 | Ethereum @ £240.14 | Litecoin @ £48.03 | Buy it online with bank transfer at"

[2,] "makes it easy to get into . Download to , , &amp; have some fun. Android: iPhone: @Bitcoin @EOS_io @NeblioTeam @Ko
modoPlatform"
[3,] "$6479.13 $279.98 Cash $535.74 $56.53 $0.34 $94.5 $141.5"

[4,] "Last litecoin price on bitstamp: $ 56.47"

[5,] "Dear &amp; swears never an with caught him Flipping $10,000"

[6,] "LTC to USD price $56.35"
```

Removal of Twitter URLs.

```
> head(LITECOIN_2)
     [,1]

[1,] "Bitcoin @ £5,410.48 | Ethereum @ £240.14 | Litecoin @ £48.03 | Buy it online with bank transfer at"

[2,] "makes it easy to get into . Download to , , &amp; have some fun. Android: iPhone: @Bitcoin @EOS_io @NeblioTeam @Ko
modoPlatform"
[3,] "$6479.13 $279.98 Cash $535.74 $56.53 $0.34 $94.5 $141.5"

[4,] "Last litecoin price on bitstamp: $ 56.47"

[5,] "Dear &amp; swears never an with caught him Flipping $10,000"

[6,] "LTC to USD price $56.35"
```

This figure below shows what the text data looks like after stop words were removed and data has now completed the data pre-processing steps ready for analysis.

```
> head(LITECOIN_2)

                                V1
1                                    bitcoin, @, £5, ,, 410, ., 48, |, ethereum, @, £240, ., 14, |, litecoin, @, £48, ., 03
, |, buy, online, bank, transfer
2 makes, easy, get, into, ., download, ,, ,, &, amp, ;, some, fun, ., android, :, iphone, :, @, bitcoin, @, eos, _, io,
@, neblioteam, @, komodoplatform
3                                                $, 6479, ., 13, $, 279, ., 98, cash, $, 535, ., 74, $, 56, ., 53, $, 0,
., 34, $, 94, ., 5, $, 141, ., 5
4                                                                                                                last, litecoin,
price, bitstamp, :, $, 56, ., 47
5                                                                                                dear, &, amp, ;, swears, never, an, caug
ht, him, flipping, $, 10, ,, 000
6
   ltc, usd, price, $, 56, ., 35
```

4.4 Summary

As mentioned above, the chapter gives an overview into the process of data pre-processing which removed hashtags, tags, URLs, twitter URLs and stop words from the text data. Whilst, having to change the order and format of the dates.

5. Analysis

5.1 Introduction

Within this chapter, a comprehensive analysis is given regarding the number of tweets mentioning each cryptocurrency and the number of video titles with their name included. Also, each cryptocurrencies price is analysed comparing its polarity and sentiment score to determine any correlation between the two.

5.2 Overall Statistical Analysis

As stated in (Aims) and objectives one of the objectives of the project was to find out which Cryptocurrency out of the four selected had the most user activity/ discussion across the social media platforms chosen. (Figure 5) shows how Bitcoin (Satoshi, Nakamoto, 2009) dominates the number of times it is mentioned via Twitter with a total of 785,279 starting on 22nd August 2018 and lasting for four days. Ripple totalled 113,419 tweets, with Ethereum in third with 89,079 and Litecoin having 11,348 tweets. According to Coinmarketcap.com (CoinMarketCap, 2018), which displays various figures around Cryptocurrencies and market shares, the order of these cryptocurrencies would be Bitcoin, Ethereum, Ripple and Litecoin in terms of market cap.

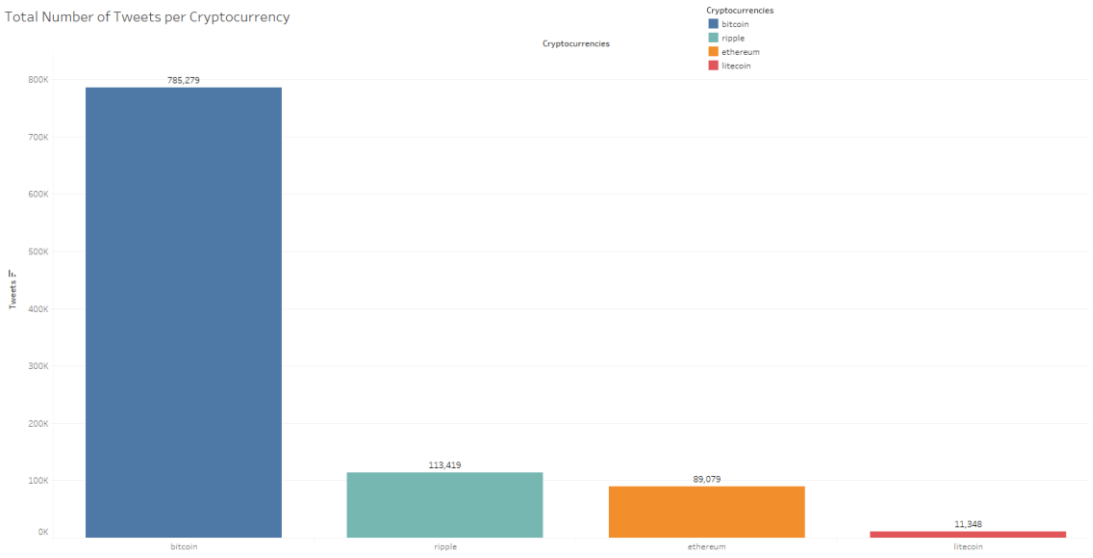


Figure 5 - Total Number of tweets per Cryptocurrency

In addition, statistical analysis was performed on the YouTube data collected, and displayed a smaller difference in totals for the cryptocurrencies as shown in (Figure 6). Bitcoin was mentioned in titles of YouTube videos 521 times, with Ethereum 352, Litecoin 335 and Ripple only having 254 mentions.

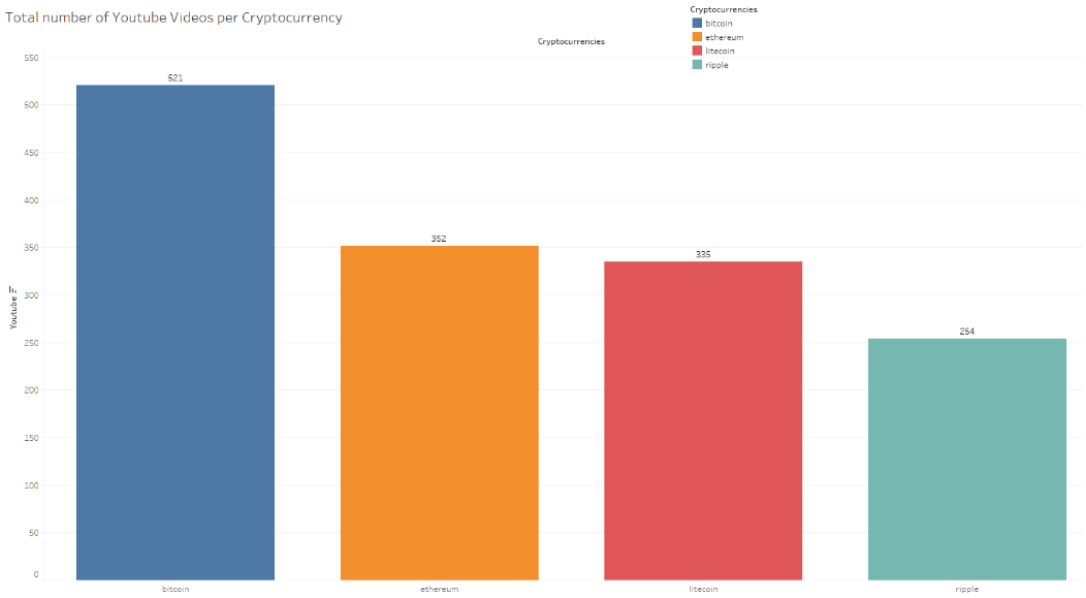


Figure 6 - Total Number of YouTube videos per Cryptocurrency

(Figure 7) shows the prices for all four cryptocurrencies across the four-day period analyzed. As shown below, there are similar trends across all four currencies, which may be due to social medias influence or other global problems with general stocks or bad news articles affecting the market all together. There is a clear correlation that prices all initially increase, until the second twelve-hour period where they all slowly decrease. Additionally, all currencies again slowly increase and provide some level of stability within their prices.

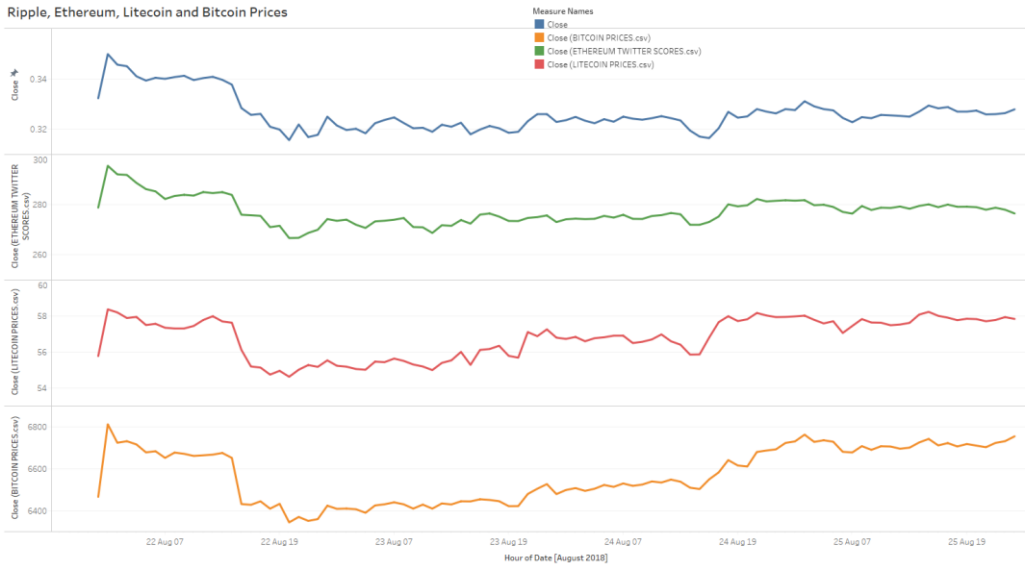


Figure 7 - Ripple, Ethereum, Litecoin and Bitcoin Prices

5.3 Bitcoin

5.3.1 Trend in Prices

Figure 8 displays the Bitcoin closing prices per hour between the dates of 22nd August 2018 and 26th August 2018. As shown, Bitcoin had one immediate rise and fall continuing to gradually

increase in its price over the following three days. Over the time period Bitcoin achieved a high of $6,813.20 and its lowest price was $6,344.20, overall increasing by 4.47%.

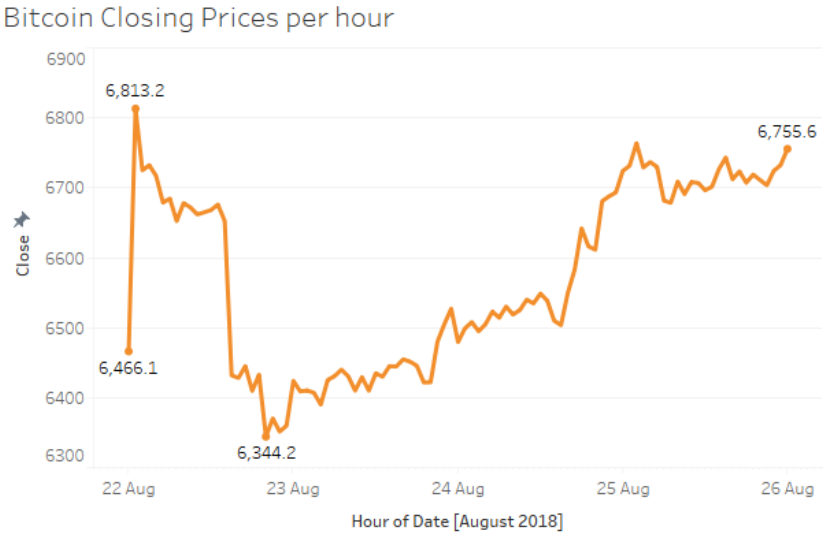


Figure 8 - Bitcoin Closing Price per hour

5.3.2 Number of Tweets

Figure 9 presents the number of tweets including the word “Bitcoin” per hour across the four-day period. The number of Bitcoin tweets shows a dramatic increase and decrease on five occasions with tweets nearly reaching 20,000 per hour on three moments. Comparing this graph to the price of Bitcoin (Figure 8), there is some correlation between the two. For example, there is a one to two-hour delay in the price fall during the first trough in the number of tweets at 13:00hrs. The price falls to $6,341.80 then decreases additionally. Furthermore, the second peak on 23rd at 17:00hrs of tweets is delayed in price increase at 20:00hrs. The second strongest similarity is on 24th where there are two increases in the number of Bitcoin tweets. This is mirrored in the price of Bitcoin with a four-hour delay of a change happening, the price between these increase does not dip as low as the number of tweets it must be said.

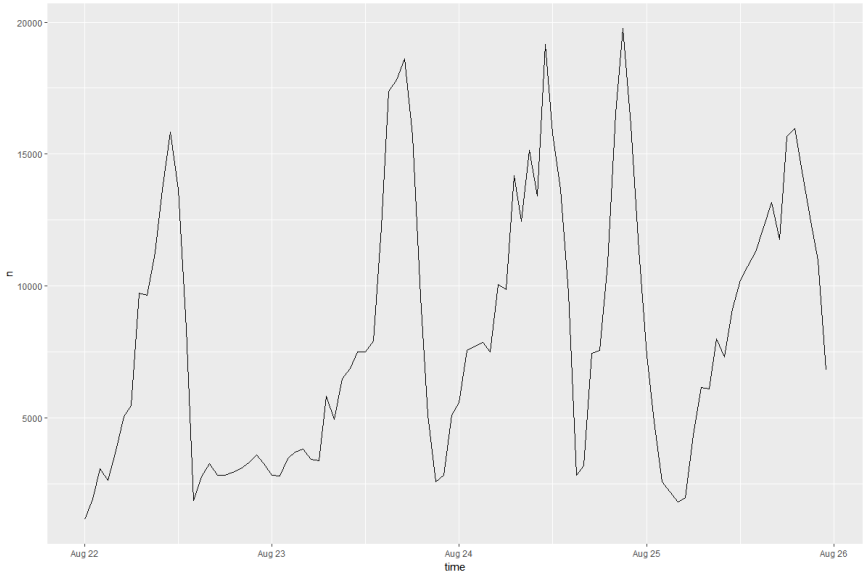


Figure 9 - Number of Bitcoin tweets per hour

5.3.3 Twitter - Polarity Score

Figure 10, shows the Bitcoin twitter polarity scores of the language used within the tweets collected. Bitcoin showed to have a 64.18% of its tweets to contain some positive language with the figure being 706,765 and the negative polarity score to be 394,384. There shows some connection with the price of Bitcoin as stated earlier the price increased over time with some falls.

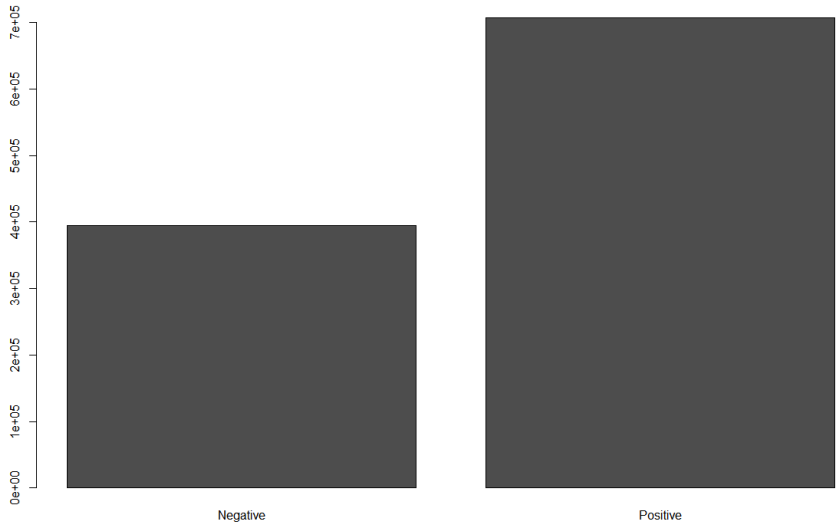


Figure 10 - Bitcoin twitter polarity score

Additionally, this data is displayed along a time frame in Figure 11. As shown below, the price has a sudden within the first twenty-four hours, this period does not seem to show a strong correlation due to positive language used outweighs the negative before and after this period. Looking at the figure, the price increases more so in the last 48 hours where there is more positive language used, some troughs are mirrored with a two-hour lag.

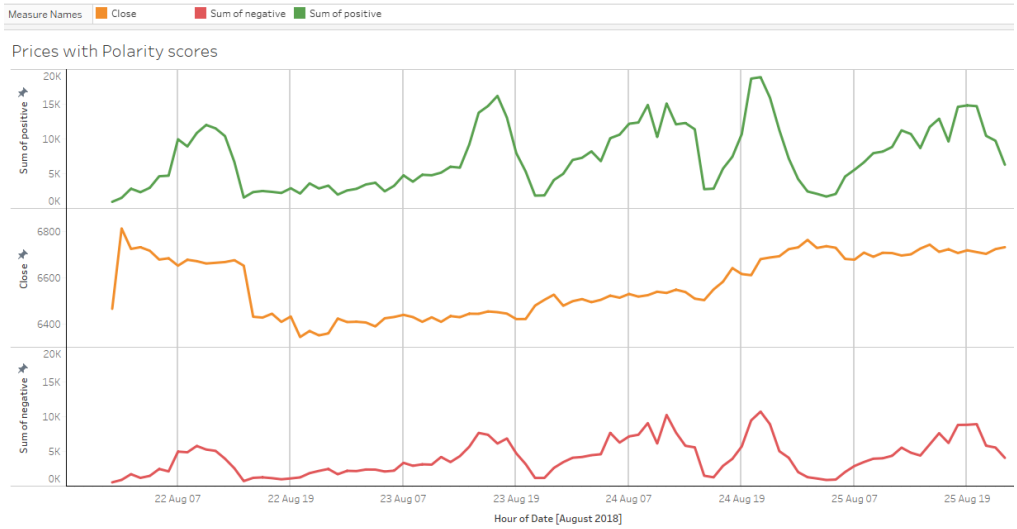


Figure 11 - Bitcoin Polarity Scores with prices

5.3.4 Twitter - Sentiment Score

Figure 12 displays the sentiment score for the language used in the Bitcoin tweets. The highest scored emotions were trust and anticipation which show a strong correlation in prices increasing overall. The emotions anger, fear and joy all show similar results when may give reason to why prices still fluctuate numerous times. Surprise is the second lowest scored emotion, which would show how individuals discussing Bitcoin via Twitter share the same knowledge that cryptocurrencies have a common occurrence of changing rapidly.

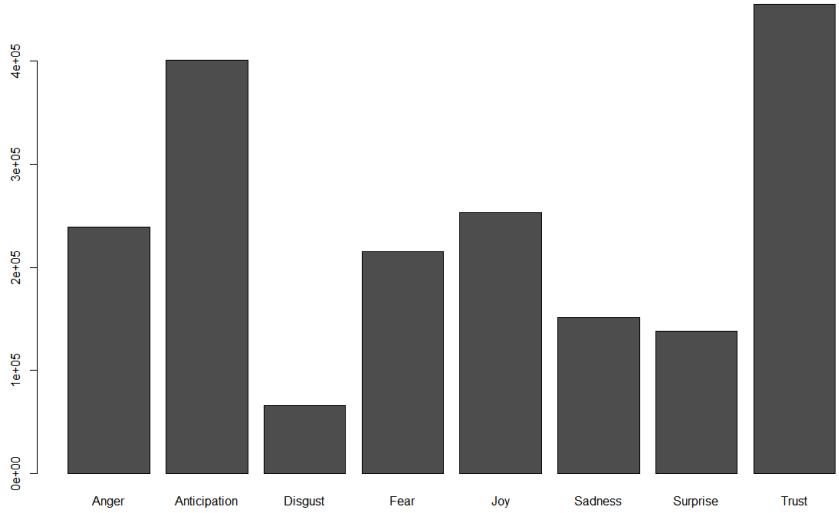


Figure 12 - Bitcoin Twitter Sentiment Score

Additionally, this data can be displayed as shown in Figure 13 below. The figure shows the anticipation sentiment score of Bitcoin with the price of Bitcoin along the selected time frame. The figure also shows that when anticipation is used within tweets and this increases significantly there is a price change two to three hours later.

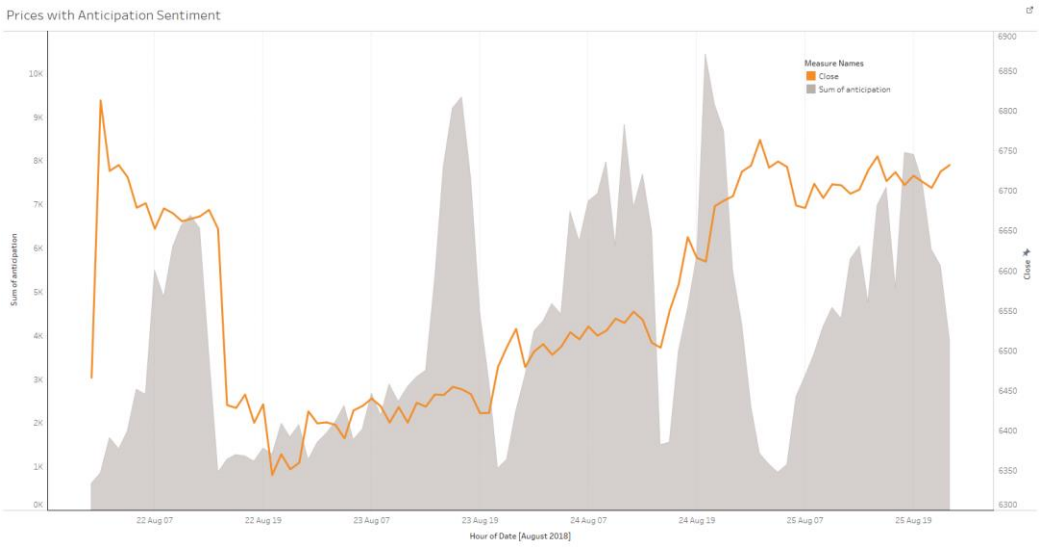


Figure 13 - Bitcoin Twitter Anticipation Sentiment with prices

Figure 14, displays the Bitcoin joy sentiment scores per hour with the price of Bitcoin along the selected time frame. This sentiment shares the same pattern as anticipation sentiment as prices increase and decrease around two to three hours after tweets containing the emotion of joy increase sustainably.

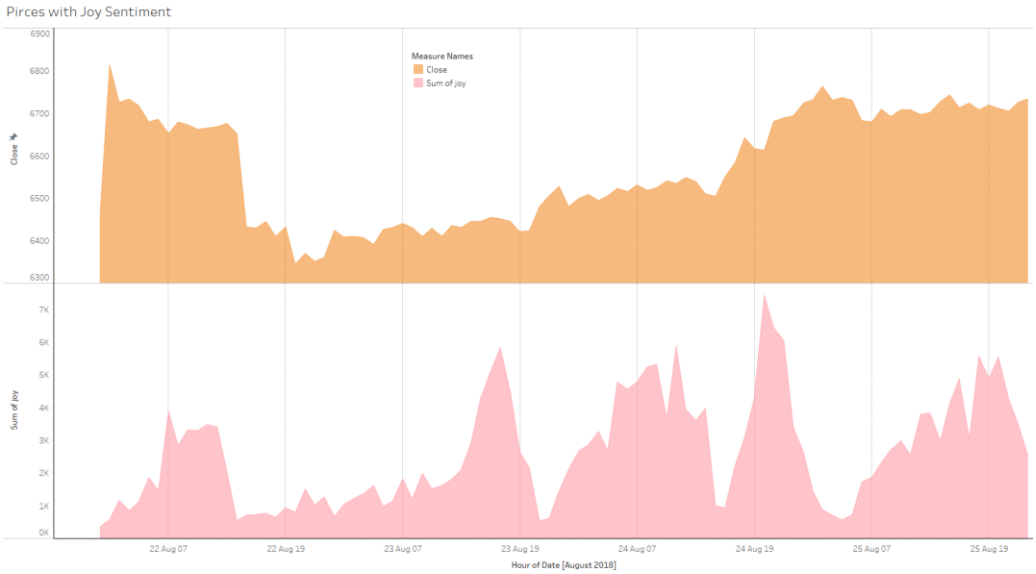


Figure 14 - Bitcoin Twitter Joy Sentiment with prices

5.3.5 Twitter – Word Cloud

The figure below, showcases the top ten words used in all Bitcoin tweets collected. As displayed, the word “bitcoin” is no surprise to be first, with “price, btc” in joint second. The words presented show that users are mainly talking about the price of Bitcoin when tweeting.

Figure 15 - Bitcoin Word cloud

5.3.6 YouTube – Number of Videos

The number of videos published via YouTube using the word “Bitcoin” in their titles per hour is shown in Figure 16. The graph shows peaks and troughs on numerous occasions and this shows some correlation with price of Bitcoin (Figure 8). From comparing the two, the number of videos created increases higher the more sudden a rise or fall happens in price. The number of videos published decreases as prices seem to reach a median. Unlike the twitter data shows, the YouTube analysis for Bitcoin shows the delay is after the price has changed. Whereas, the price has a delayed time depending on the number of tweets.

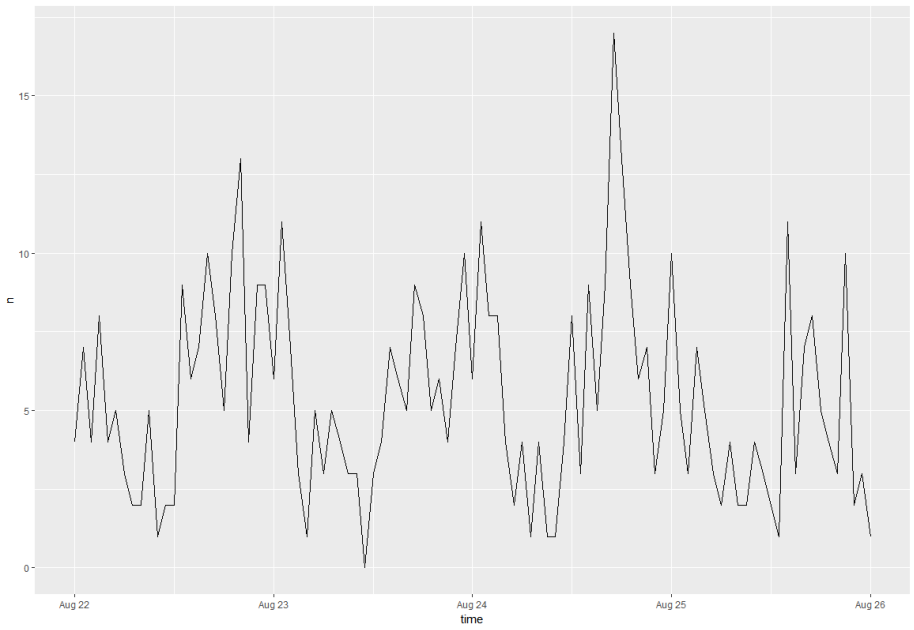


Figure 16 - Number of Bitcoin YouTube videos per hour

5.3.7 YouTube – Polarity Score

The below figure, shows the polarity score of Bitcoin YouTube video titles. As shown, the positive score of polarity had only 56.11% of the total. The exact score was positive with 156 and negative of 122. This shows again, some relevant connection with the price of Bitcoin due to prices increased over the time period.

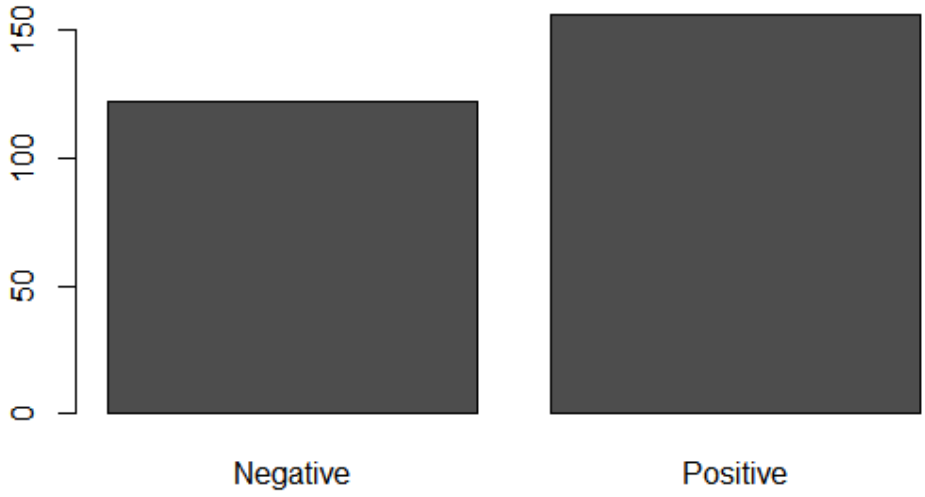


Figure 17 - Bitcoin YouTube Polarity score

5.3.8 YouTube – Sentiment Score

The YouTube data was explored further, the emotional sentiment was analysed of the YouTube titles containing Bitcoin as is shown in Figure 18. As displayed, the YouTube data shows similarity with the Twitter sentiment (Figure 12) both showing trust and anticipation to be ranked as the top two. This time the emotion sadness has increased and with the emotions fear and anger. The results of this sentiment show some correlation with the number of peaks and troughs within the price of Bitcoin.

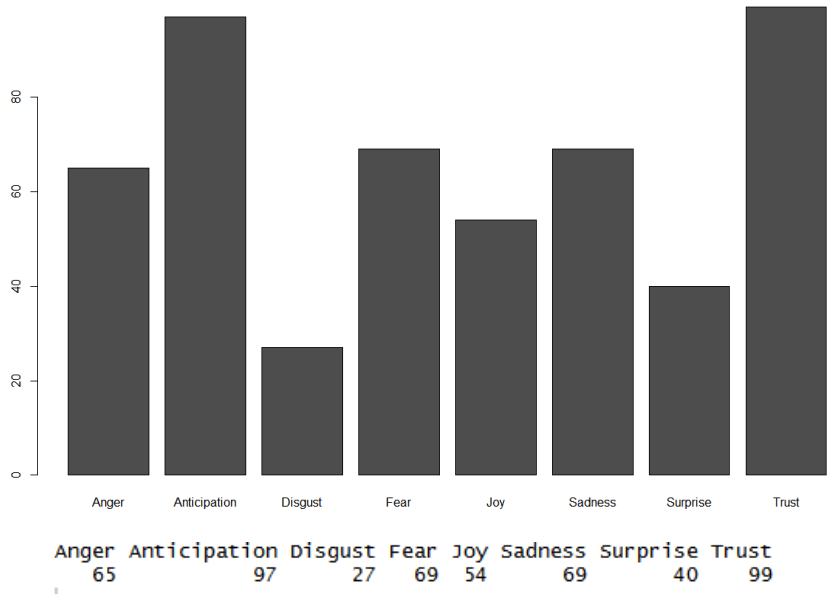


Figure 18 - Bitcoin YouTube Sentiment Score

5.4 Ripple

5.4.1 Trend in Prices

Figure 19, shows the closing prices for Ripple every hour between 22nd August and 26th August. The figure details how the price of Ripple increased to its highest value over the four days reaching $0.3498. The price of Ripple decreased over the next twenty hours, this time reaching a low of $0.3157. After this, Ripple displayed a variety of peaks and troughs throughout the final seventy-two hours. Overall, the price of Ripple decreased by 1.31%, which is not an enormous amount compared to some other cryptocurrencies in the past.

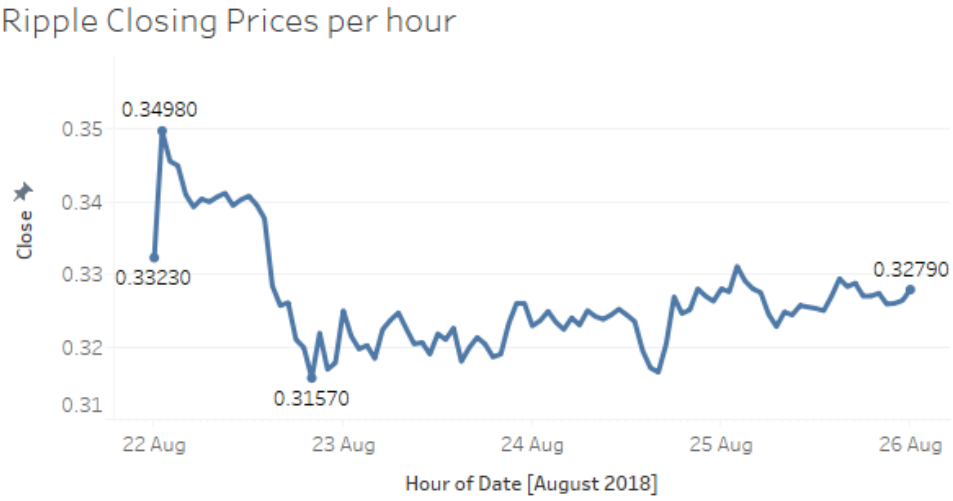


Figure 19 - Ripple closing prices per hour

5.4.2 Number of Tweets

Figure 20 displays the number of tweets mentioning the word Ripple per hour. As shown there are three major peaks and troughs during the time period with a number of small peaks and troughs. Three substantial increases happen with the first peaks at 12:00hrs on 23rd August, the following two happen 24 hours after one another. Within 12 hours after the increases, the price falls to its original price before the increase. Comparing this graph to (Figure 19), there a weak correlation between the two. Some peaks and troughs are mirrored as the first twelve hours show both an increase and decrease. Though it must be mentioned that peaks and troughs are not as violent in one as the other. For example, on 23rd August 12:00hrs the number of mentions increases this is not shown in prices until 22:00hrs with a small increase. Again, on 24th August at 04:00hrs tweets increase until 13:00hrs, prices start to increase at 16:00hrs. Arguably showing there is around a 12-hour delay from tweets to prices, with some correlation.

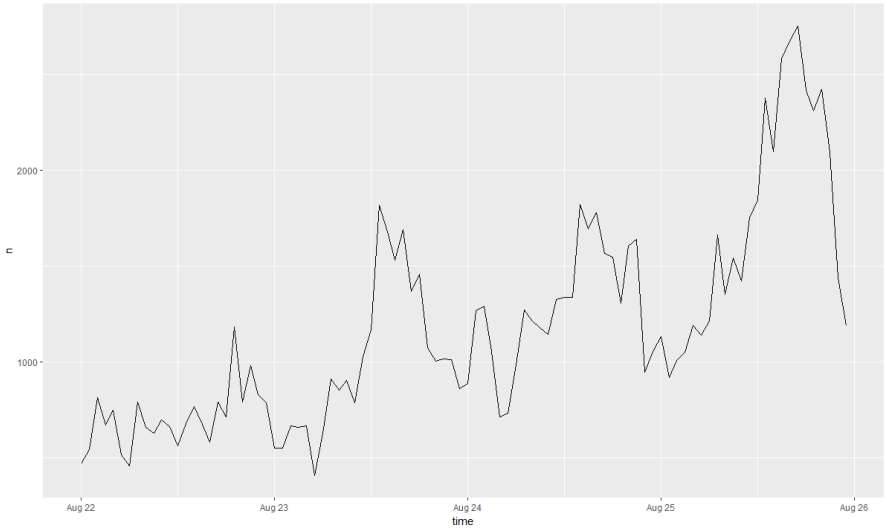


Figure 20 - Number of Ripple Tweets per hour

5.4.3 Twitter - Polarity Score

Figure 21 presents the Twitter polarity score of Ripple, showing how many tweets contained positive and negative language. The results show the negative score to be 71,540 and positive language used to be 166,440 meaning the overall language used by Ripple users was 69.9% positive. Comparing this to the price of Ripple, it does not show an initial correlation as such due to a price decrease overall of 1.31%. Though, taking from these results it could mean that the level of positive language and discussion via Twitter needs to be a majority of positive to keep prices at a stable rate.

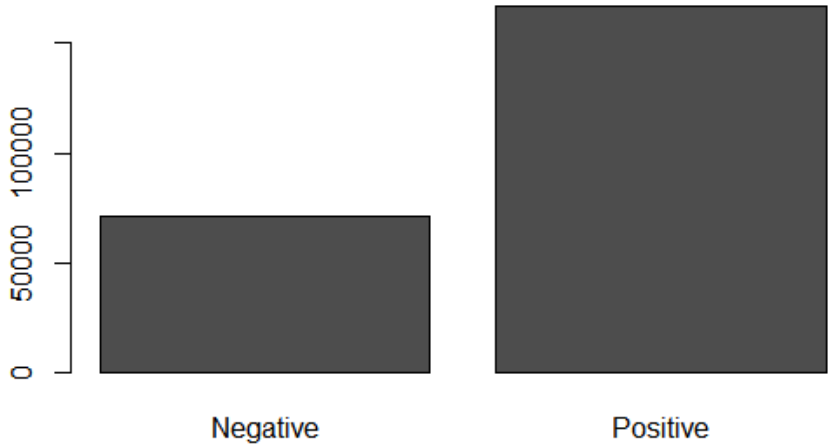


Figure 21 - Ripples Twitter Polarity Score

Furthermore, Figure 22 showcases this data more with a time series analysis of the polarity scores against the price of Ripple. As laid out, on two occassions specifically on 22nd and 24th 19:00hrs, prices show a decrease the number of positive language used in tweets increases by a significant margin with prices continuing after at a more steady rate.

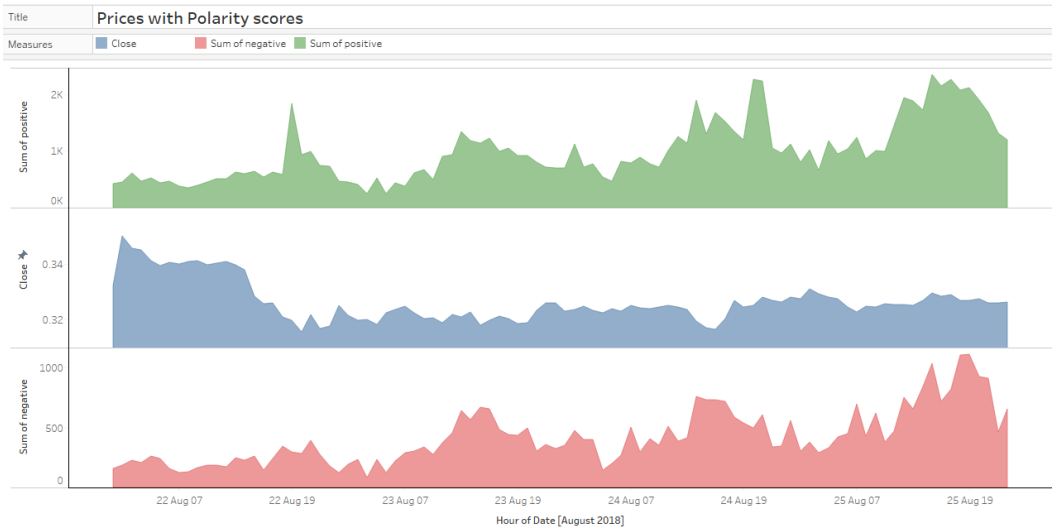


Figure 22 - Ripples Twitter Polarity Scores with prices

5.4.4 Twitter - Sentiment Score

The language of Ripple tweets was explored further, Figure 23 shows the emotional sentiment scores for this. As shown, the emotions trust and anticipation are the highest scored by an expressive margin, with anger, fear and joy all sharing relatively similar scores. These results seem fairly accurate towards the language used by twitter users towards Ripple. As, shown in the price of Ripple, the cryptocurrency community still show a level of trust and hope for prices even though the smallest decrease.

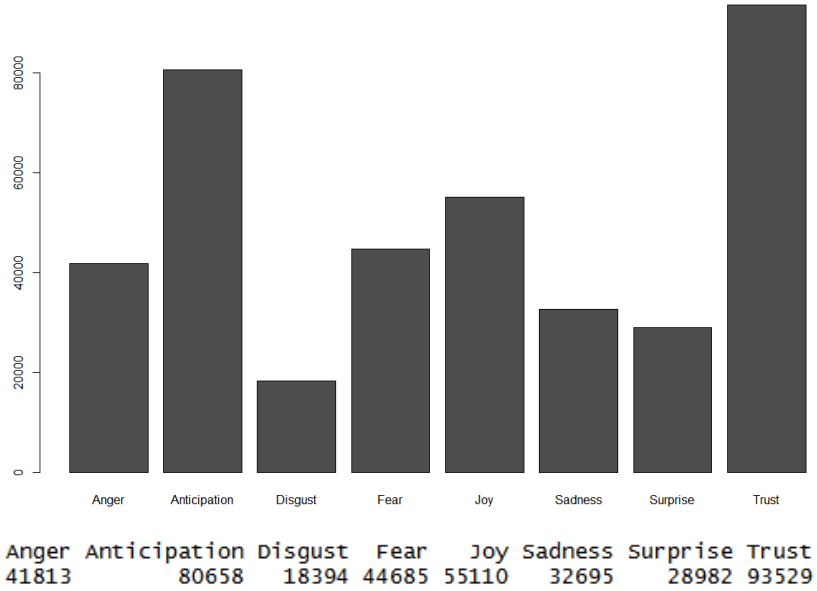


Figure 23 - Ripples Sentiment Score for Twitter

Arguably, Figure 24 appears to show a correlation between the sentiment score of fear and the prices for Ripple. Fear shows a 11.2% share of the total sentiment score. As displayed below, when the price Ripple decreases by a generous amount, the number of tweets using language related to fear increases by a serious measure. This means users using Twitter still show some level of fear towards price devaluation.

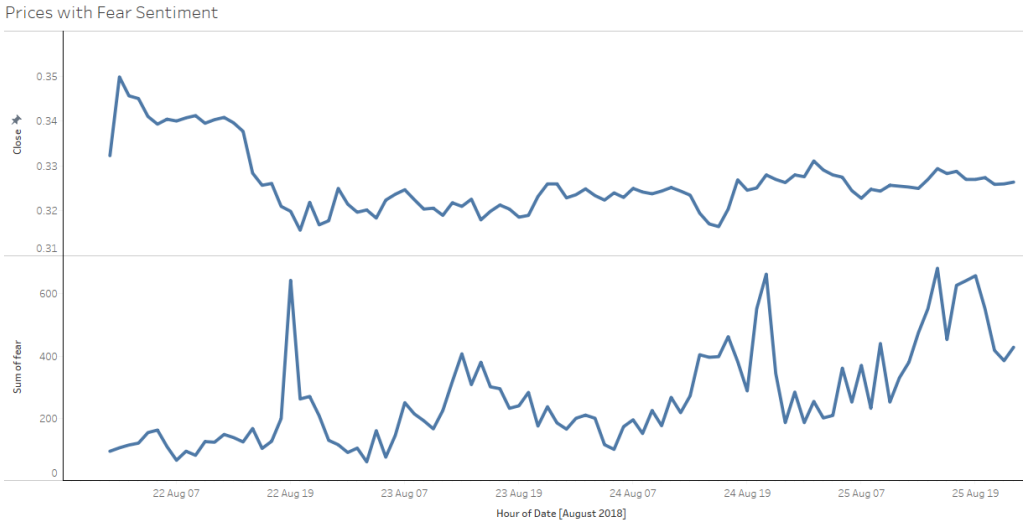


Figure 24 - Ripples prices with Fear Sentiment

5.4.5 YouTube – Number of Videos

Figure 25, displays the number of YouTube videos which use the word “Ripple” in their title. Comparing this to (Figure 19), this graph shows a correlation in number of peaks and troughs. It would seem that the number of YouTube videos created does not have a knock-on effect on the price, yet it shows a mirrored affect especially in how fierce the peaks and troughs are. When the price of Ripple both increases or decreases the number of videos rises.

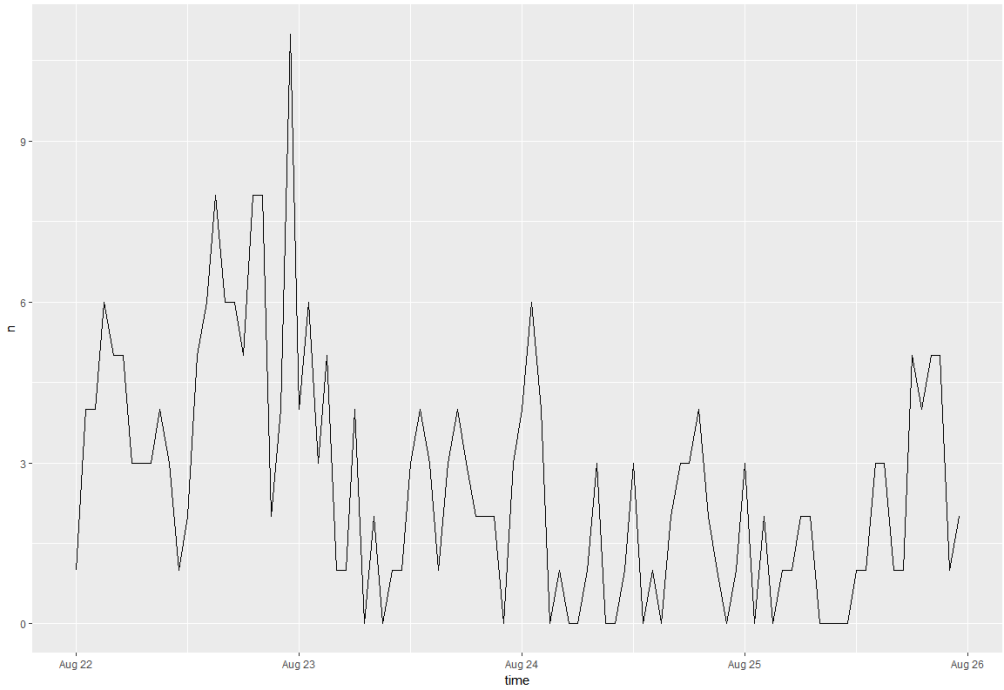


Figure 25 - Ripples number of YouTube videos per hour

5.4.6 YouTube – Polarity Score

Figure 26, shows the polarity score of YouTube video titles displaying the word Ripple within them. As shown the polarity scores are close with the negative scores to be 66 and positive to be 82. From the results, it would show some correlation with prices are they show an even number of peaks and troughs. Though, analysing the sentiment of the video titles may be slightly inaccurate to the content of the video.

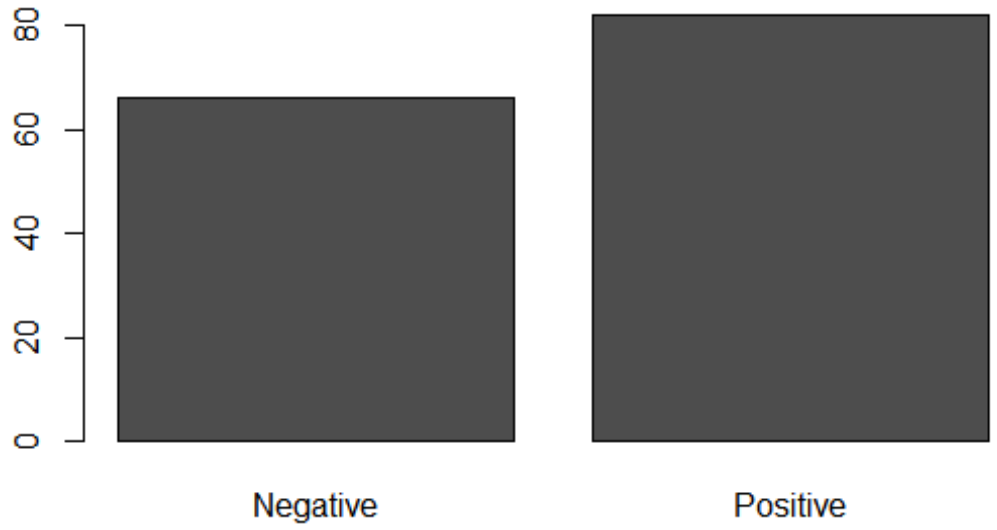


Figure 26 - Ripples YouTube Polarity Score

5.4.7 YouTube – Sentiment Score

Figure 27, shows the sentiment score of the YouTube titles containing the word Ripple. Similar results are presented in comparison with Twitter (Figure 23), with anticipation and trust being

the top two emotions. This time fear is presented more in the language used instead of joy but are closely followed by anger too.

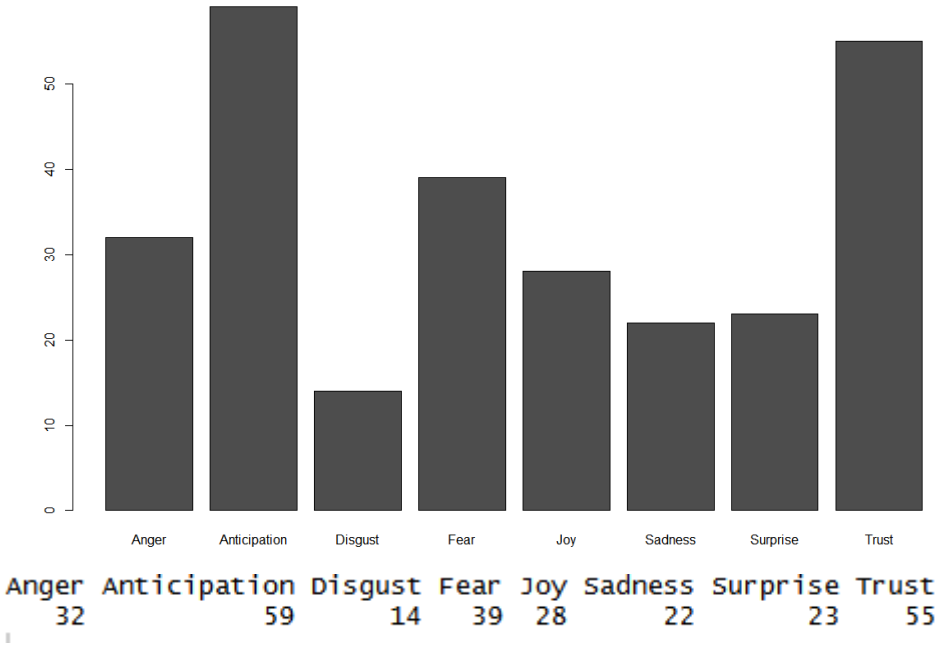


Figure 27 - Ripples YouTube Sentiment Score

5.5 Litecoin

5.5.1 Trend in Prices

Figure 28, displays the closing price of Litecoin per hour over four days. The figure shows a variety of fluctuations across the time scale and sees the price increase to $58.370 with its lowest being at $54.630, this happened within the first 24 hours. The overall price increased by 3.69% and showed to slowly increase between the last 72 hours.

Figure 28 - Litecoin Closing Prices per hour

5.5.2 Number of Tweets

The figure below shows the number of tweets per hour which included the word “Litecoin” during the four-day-period. The initial conclusion from this graph is there are two major peaks and troughs one on 22nd August 19:00hrs and the other on 25th August 07:00hrs. Comparing this to Figure 28 above, these show some correlation specifically the first time the number of tweets see to peak the price of Litecoin seems to drop showing a mirror effect in terms of reacting. The second peak does not show a strong connection to the change in prices as much. Additionally, the number of tweets per hour does increase over time which is similar to the prices increasing after the four days too.

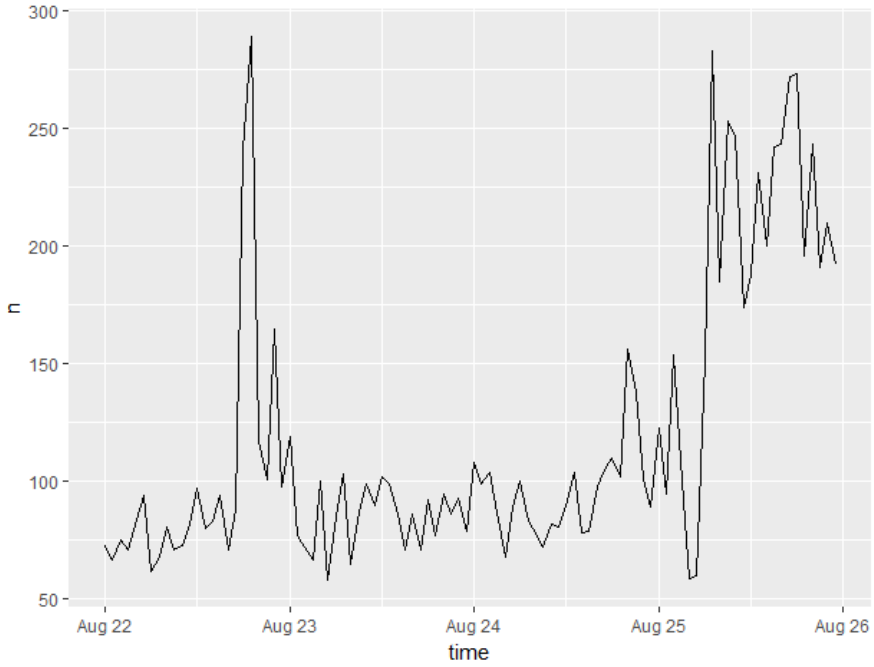


Figure 29 - Number of Litecoin Tweets per hour

5.5.3 Twitter - Polarity Score

The figure below shows the polarity score of Litecoin tweets. Stating the obvious, the positive score of the tweets has a vast clearance over the negative scoring, with 77.79% of the total scoring. The positive polarity of the tweets scored 9,135 whereas negative only scored 2,608. This does show a correlation with the price difference in Litecoin overall as it increased which shows the tweets using positive language could have an effect on the price of Litecoin.

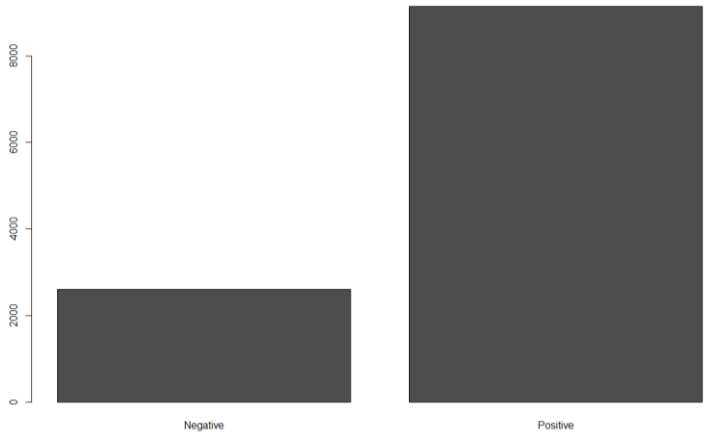


Figure 30 - Litecoin Twitter Polarity Score

5.5.4 Twitter - Sentiment Score

Continuing on, the figure below displays the sentiment score for Litecoin tweets showcasing eight different emotions. From the off, it is clear to state that the emotions trust and anticipation are the top two detected in the tweets, which helps to possibly explain why prices change continually. Likewise, fear, joy and anger are very close in their detection levels sharing similar scores.

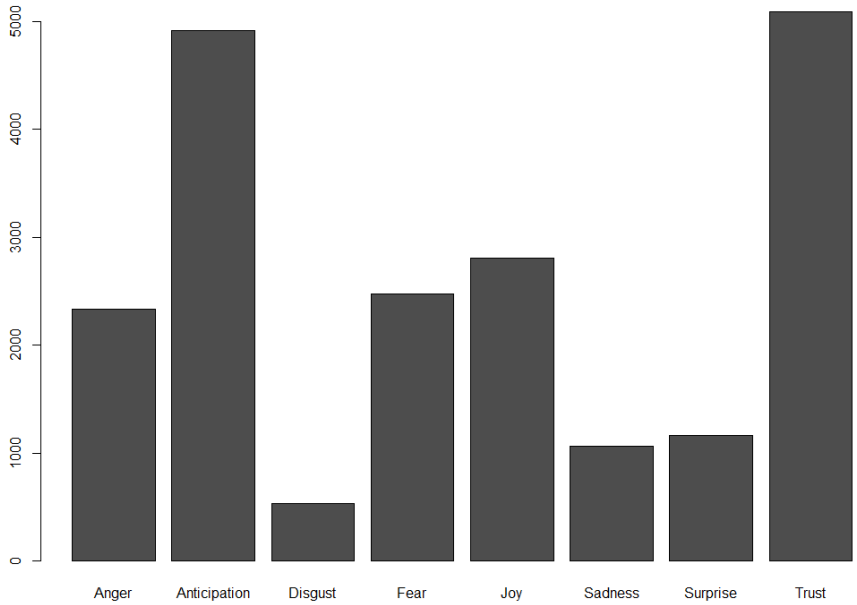


Figure 31 - Litecoin Twitter Sentiment Score

Furthermore, the figure below the connection below sentiment detection of anticipation within Litecoin tweets and the price of Litecoin. As shown, there is a delayed reaction this time from the number of tweets with anticipation as it shows a rise after a significant decrease in the value of Litecoin within the first 24 hours. There is a stronger link between the smaller peaks and troughs throughout the timeframe.

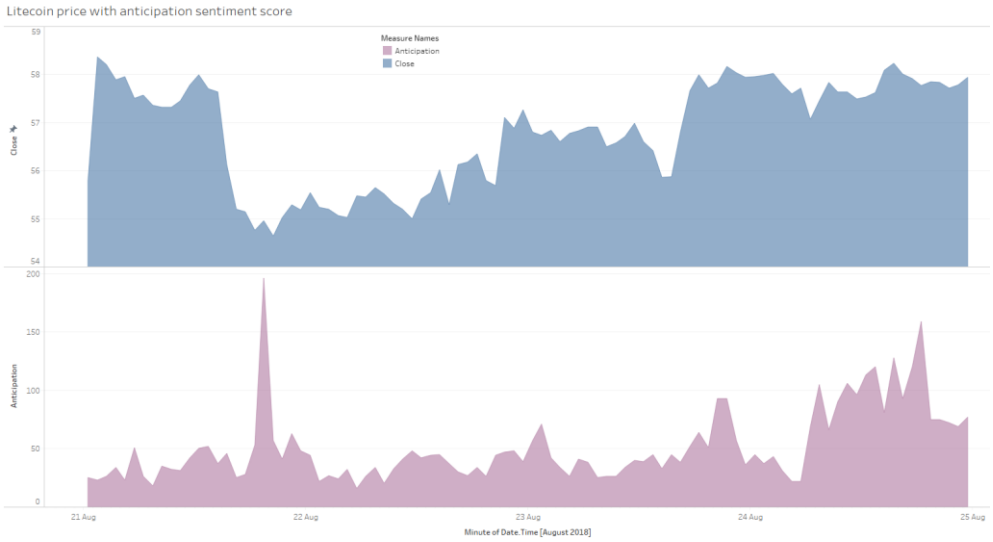


Figure 32 - Litecoin price with anticipation sentiment score

5.5.5 Twitter – Word Cloud

The word cloud below shows the top ten words used within the tweets all containing the word "Litecoin". Interesting, both "Ethereum and Bitcoin" are both used simultaneously with the word when users' tweets. Bitcoin even is showed to be the second most popular word used. This may be due to individuals comparing the other cryptocurrencies when Litecoin fluctuates in price.

Figure 33 - Litecoin Word Cloud

5.5.6 YouTube – Number of Videos

Figure 34, displays the number of YouTube videos per hour which contain the word "Litecoin" in their title during the selected time scale. The graph shows a variety of peaks and troughs throughout, which the number changing constantly. Comparing this with the price of Litecoin (Figure 28), the number of YouTube videos published shows some correlation as the first major peak in price happens just after a large peak in videos published. Additionally, the second major peak in price is mirrored in the number of videos created, the small peaks and troughs are not similarly connected.

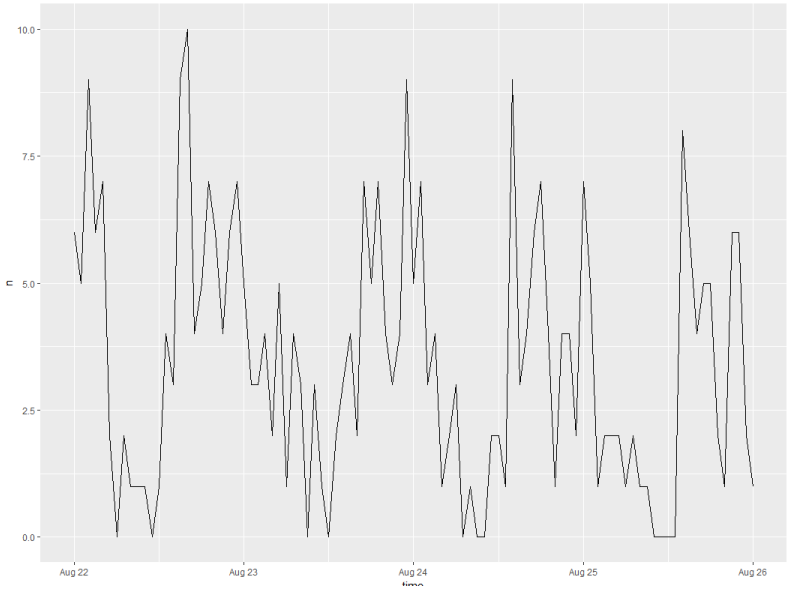


Figure 34 - Number of Litecoin YouTube videos per hour

5.5.7YouTube – Polarity Score

The figure below shows the polarity score of Litecoin YouTube video titles created. As shown, the scores of both are very tight with the positive polarity sharing only 50.45% of the total detected. The positive polarity detected was 112 and negative scoring 110. This does not show a correlation compared with prices of Litecoin overall as again prices did increase slightly. Though this may show a relationship between the small increases and decreases in prices.

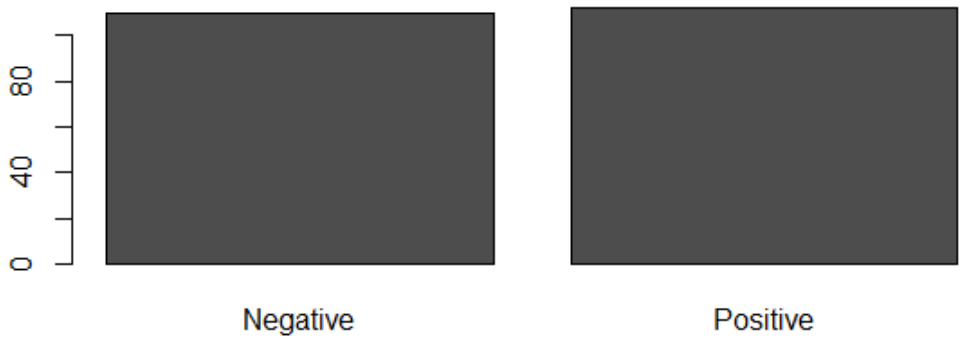


Figure 35 - Litecoin YouTube Polarity Score

5.5.8 YouTube – Sentiment Score

The figure below shows the sentiment scores of Litecoin YouTube video titles. The emotion fear displays the highest figure with trust 11 behind. Anticipation, anger and sadness are share very similar scores. These sentiment scores do not show a similar compared with Litecoin twitter sentiment scores (Figure 31) and again does not show a clear correlation towards Litecoin prices. Though the results may in addition explain why there is constant uncertainty and change within daily prices of Litecoin altogether.

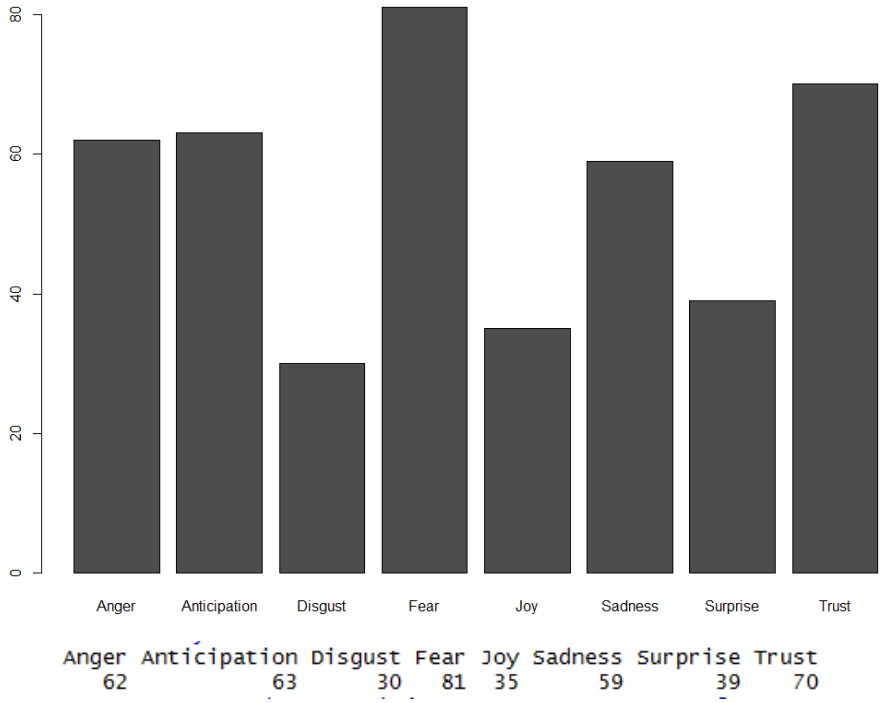


Figure 36 - Litecoin YouTube Sentiment Score

5.6 Ethereum

5.6.1 Trend in Prices

Over the four-day period, Ethereum prices decreased by $2.33 which is not an enormous amount overall. Though, Figure 37 displays prices rising to $295.50 within an hour and prices dropping as low as $266.71 nineteen hours after which slowly increased. Throughout the four days prices fluctuated some more severe than others, hopefully looking further into the social media data this may provide some answers into why this happened.

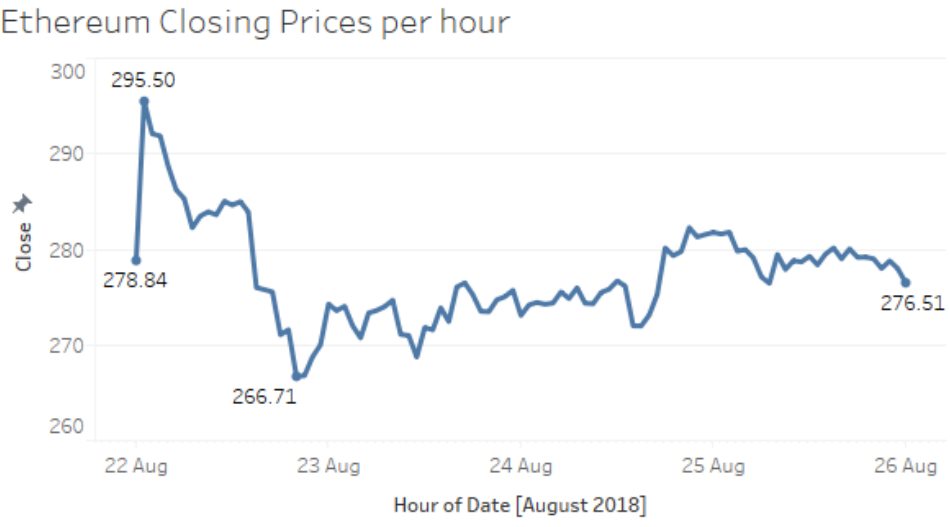


Figure 37 - Ethereum Closing Prices per hour

5.6.2 Number of Tweets

Figure 38 shows the number of tweets per hour over the four days, as shown below there is a substantial increase the number mentions of the word “Ethereum” over the Twitter platform. Mentioning’s of the word rose especially on the 25th August reaching highs of nearly 3,000 tweets per hour. Comparing both figures, prices decreased overall and the conversation of Ethereum has risen in total. Looking further into the peaks and troughs, the initial increase in price is not mirrored as the number of tweets display a minimal increase. It would seem as conversation increasing around Ethereum its price gains some stability with a reduced number of such extreme price changes. The minor peaks and troughs within both graphs show more similarity, with some delay proven.

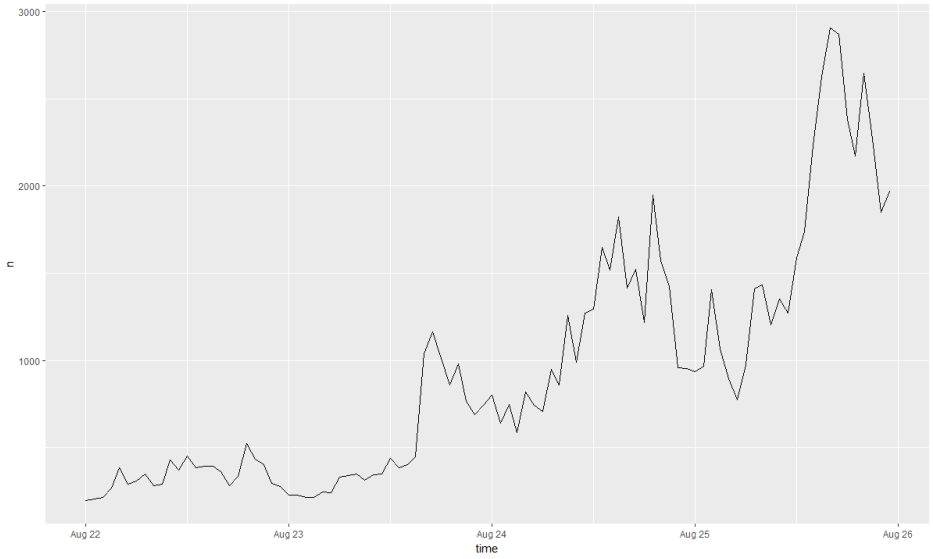


Figure 38 - Number of Ethereum Tweets per hour

5.6.3 Twitter - Polarity Score

Figure 39, provides an understanding of the language used when users mention Ethereum in their tweets. As discussed in (Methodology), polarity scores give an indication of whether conversation is positive or negative. Ethereum had a positive score of 89,439 and negative of 23,449 resulting in 79.2% of the mentioning being classed as positive. (Literature Review), discusses lexicon-based methods can have difficulty in classifying words inaccurately so this must be reminded looking at the positive score being greater.

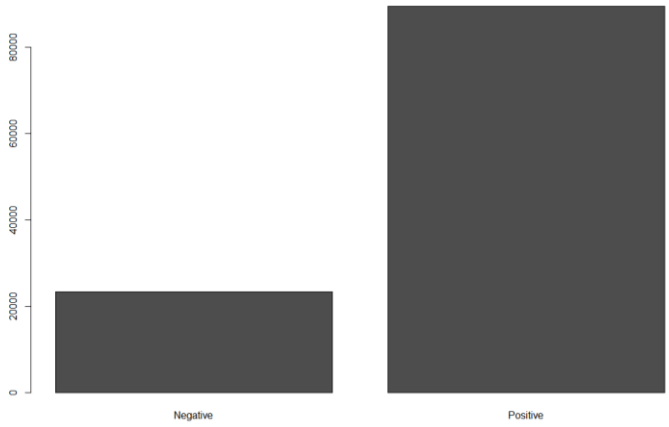


Figure 39 - Ethereum Polarity Score

Whilst analyzing Figure 40, it shows that there is no clear similarity in negative polarity scores towards prices within the Ethereum community of Twitter users. This is because there is no rise of negative discussion before or after prices decreases over the first twenty-four-hour period. Whilst on the other hand, the number of positive mentions vastly increases which may have an impact on prices stabilizing between 23rd and 26th.

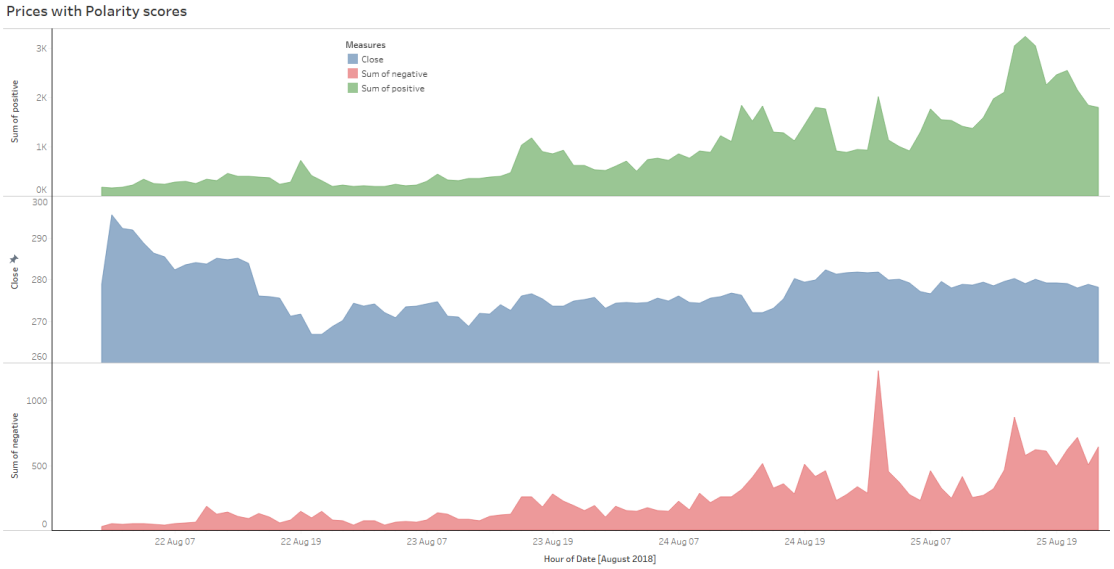


Figure 40 - Ethereum Polarity scores with prices along timeline

5.6.4 Twitter - Sentiment Score

Figure 41 displays the emotion score of the text used when mentioning Ethereum via Twitter. Trust and anticipation are shown to be the two highest scorers which may explain why there was not an increase in activity when prices dropped in the first twenty-four-hour period. Users talking about Ethereum show to be little surprised in their language used when prices fluctuate so often. Joy, fear and anger are all similarly balanced which is understandable as again the price only decreased by 0.83%.

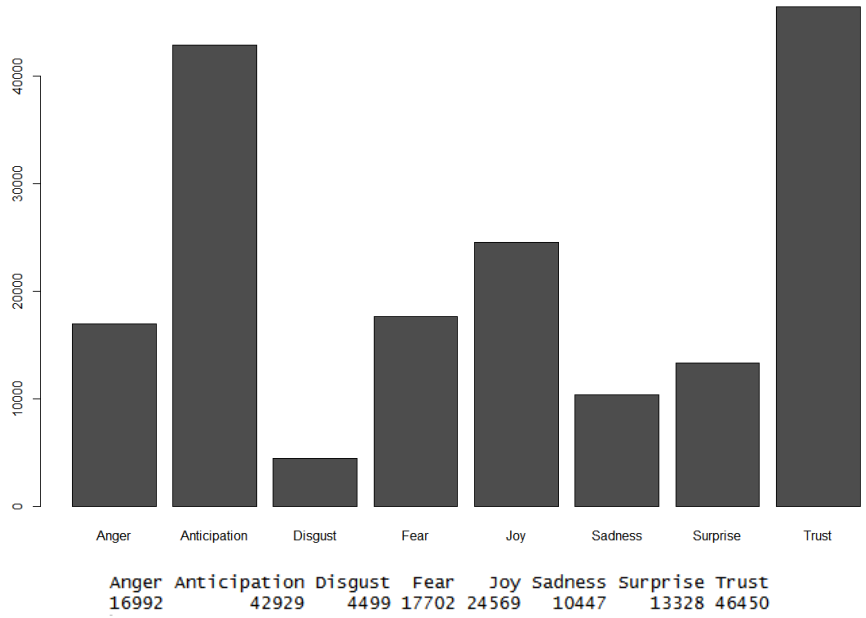


Figure 41 - Ethereum Sentiment Score

5.6.5 YouTube – Number of Videos

In terms of YouTube analysis for Ethereum, one of the variables measured was the number of YouTube videos containing the word Ethereum in their title. This result can be shown in Figure 42, which shows a number of peaks and troughs throughout the four days. The number of videos being released show a more positive correlation to Ethereum prices shown in Figure 37 compared with the number of tweets being made about Ethereum. Initially, there is a mirrored approach as prices increases the number of videos made does too. More specifically in the first twelve hours both prices and videos start to decrease. Interestingly though, as prices start to descend below 22nd August starting out price of $278.84, the number of videos increasing clearly reacting to the sudden fall in prices between the times of 13:00 and 20:00. Overall, the number of videos published show several changes in the volume of videos being posted though there is a small decrease from compared with the starting point of 00 hours of 22nd August. This is a similar pattern of the Ethereum close hourly prices.

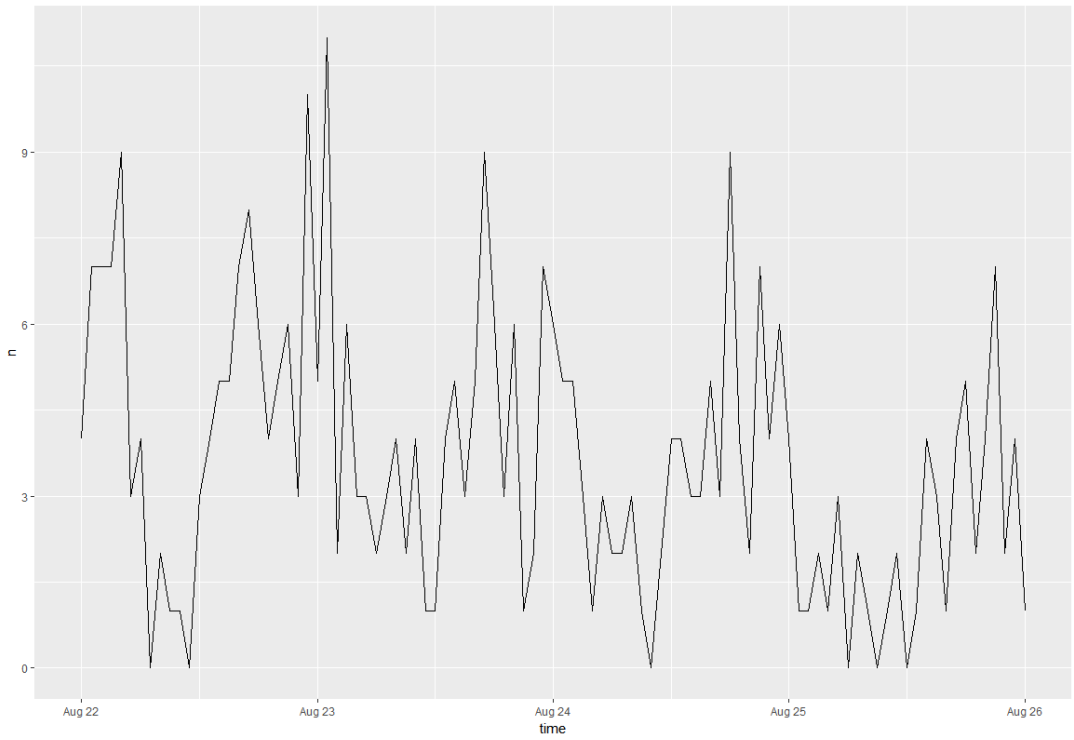


Figure 42 - Number of Ethereum YouTube videos

5.6.6 YouTube – Polarity Score

As shown in Figure 43, the polarity scores of YouTube videos towards Ethereum are very close. Ethereum scored a negative score of 93 and positive score of 99. This is completely different from the twitter polarity scores and has more similarity with prices due to small decrease. Though it must be stated that this polarity was measured on the title text of the video, whilst twitter allows a user to be more expressive in a “tweet”.

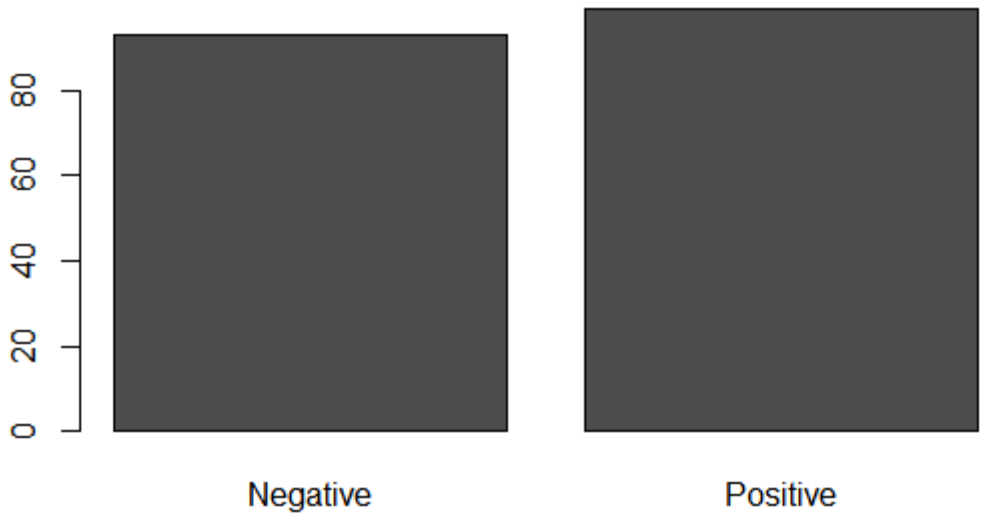


Figure 43 - Ethereum YouTube Polarity Scores

5.6.7 YouTube – Sentiment Score

Figure 44 shows the sentiment score achieved by the video titles created relating to Ethereum. Similar to the twitter results both trust and anticipation are top, this time sadness is shown to be one of the highest emotions detected along with fear and anger. This may be explained by users of YouTube creating titles to catch other users' attention to the rise or fall in prices for Ethereum.

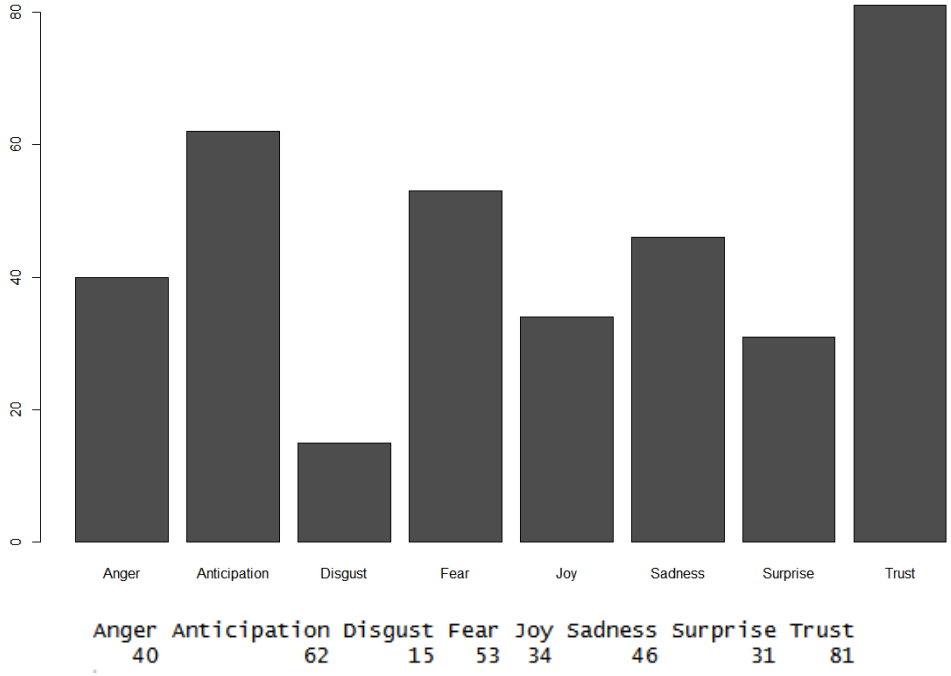


Figure 44 - Ethereum YouTube Sentiment score

5.7 Summary

From this chapter, it is clear to state that prices for Bitcoin and Litecoin increased whilst Ethereum and Ripples decreased across the time frame. In addition, Bitcoin and Litecoin shows the strongest correlation out the four currencies between their price and social media discussion.

6. Issues and Challenges

6.1 Introduction

This short chapter outlines the issues with data quality throughout this project and the limitations it has in terms of sentiment accuracy, API accuracy and the issues of focusing on one exchange. The chapter also outlines the general problems faced in the project.

6.2 Data Quality

The first data quality issue to discuss is sentiment analysis, already this issue has been raised in (Accuracy of Sentiment Analysis). The use of sentimental analysis cannot fully determine whether social media has a one-hundred percent effect on Cryptocurrency prices. This is due to contextual understanding (Brandwatch, 2015) (i.e. Sarcasm) when used in posts on social media. For example, a tweet might say “Bitcoin’s prices have dropped. Brilliant!”. Using the proposed method of analysis, this would categorise this tweet under ‘Brilliant – positive – Bitcoin’ which would be wrong. The same issue would be for biased tweets, this is something which cannot be completely resolved. It would take thousands of hours to manually go through and check the accuracy of the sentiment scores. Biased tweets may be seen as more of a popularity towards a currency. This is something which would need to be considered when considering the results of the analysis. To help prevent this, as debated a pre-packaged sentiment tool was used when analysing the data.

Another, data quality problem which needs to be considered is this data is only focusing on one exchange over a four-day-period. The exchange was chosen due to its popularity and also that United State Dollars is one of the most common currencies to exchange into cryptocurrencies.

Additionally, the Twitter API (Twitter Developers. 2018) has its restrictions and data quality issues too. The standard search function states on the developers’ page, that it focuses more on relevance to the search and not completeness. Stating some data may be missing and the standard API is limited to a seven-day search period.

6.3 General Project Challenges

As explained above, one of the main challenges within the project was to deal with issues regarding data collection and how to tackle this with the limited resources the project had. The main challenge during the project occurred when the Twitter data was lost, resulting in re-collecting the data, cleansings the data and processing the data. This was due to the R studio (RStudio, 2018) workspace becoming unresponsive when trying to save the workspace with the new collected social media data. This resulted in the analysed time frame being shortened in order for the next stages to be completed in time for the project to succeed.

6.4 Summary

As discussed, the main issues regarding data quality was the API restricted accessibility to gathering all tweets and focused more on relevance. In addition, accuracy of the sentiment analysis could not be checked manually due to this would be a near impossible task given the timeframe of the project.

7. Evaluation

7.1 Introduction

This final chapter, outlines a summary of the results from the analysis, whilst detailing the limitations of this work specifying areas for future work.

7.2 Conclusion

Overall, the conclusion made from the analysis is that the Cryptocurrency with the highest popularity of discussion across both Twitter (Twitter Developers. 2018) and YouTube (Google Developers. 2018) is Bitcoin (Satoshi, Nakamoto, 2009). As shown in the analysis section, 785,279 Bitcoin tweets and 521 YouTube videos were collected starting on 22nd August 2018 across a four-day period. This is reflected in the price of Bitcoin (Satoshi, Nakamoto, 2009) it is the highest priced Cryptocurrency currently on the market. Arguably, the price of a cryptocurrency does not necessarily reflect the amount of user activity as Ripple displayed. Ripple was placed second in the total number of tweets being 113,419, being ranked third out of the four currencies and presents the lowest price out of all four by a fair majority.

In terms of polarity scores, Bitcoin, Ripple, Ethereum and Litecoin all detected a greater positive score across both Twitter and YouTube. Bitcoins highest emotion calculated across both social platforms was Trust, with Ripple detecting Trust for Twitter (Twitter Developers. 2018) and Anticipation for YouTube. Litecoin scored Trust at its highest emotion through tweets and fear in YouTube video titles. Finally, Ethereum like Bitcoin detected Trust to be its highest emotion across both platforms.

Furthermore, the results show that Bitcoin and Litecoin show some level of correlation between social media's discussion and their prices. This observation has been made due to both prices increased over the four days and the amount of discussion increased and its language used was positive. Ethereum and Ripple showed to have a weak correlation between their social medias influence and prices due to both prices decreased with discussions increasing and again both cryptocurrencies showing a positive polarity score.

Comparing these results to others in the field, Phillips and Gorse (2017) specific how using social media data in their case Reddit (reddit, 2018), has provided "some empirical evidence that bubbles mirror the social epidemic-like spread of an investment idea" (Phillips and Gorse, 2017, pp 7). Also, Kim, et al. (2016), discovered when trying to predict future cryptocurrency prices using social media that Bitcoin fluctuations in price showed a strong connection more with positive comments. Whereas, other cryptocurrencies showed negative comments to present the strongest correlation. Kim, et al. (2016) additionally, detecting an 8% accuracy gap between their price predication using social media data and the actual figure.

7.3 Limitations and Future Work

As briefly discussed one of the limitations to this piece of work, would be centred around the accuracy of the sentiment analysis carried out. This is not specifically aimed at just this paper but the accuracy of measuring sentiment across text mining altogether. This would definitely be an improvement to measure the accuracy of the sentiment carried out in this paper and additionally, improving the method of sentiment within the social science community.

Another limitation of this work could be number of days analysed as this is only four days. Though as explained the original was damaged and the standard Twitter API (Twitter Developers. 2018) only allows seven days to be collected and not all tweets are collected. So, another reference for future work would be to invest into purchasing the premium or enterprise search API Twitter (Twitter Developers. 2018) offer and use this to collect tweets over a longer period of time to see if there is any correlation between the two variables. This would only increase the number of tweets, which helps improve detecting the accuracy levels from the method of sentiment analysis.

Lastly, more research could be carried out into other cryptocurrencies and discovering whether they have any correlation between social media discussions and their prices. This research focused on Bitcoin, Ethereum, Ripple and Litecoin as the focus of others research is towards Bitcoin alone.

7.4 Project Success

In terms of whether the project was successful, it was stated at the start of the project that all objectives must have been met. Referring to the (Objectives), number three objective stating the use of Apache Spark (Apache Spark, 2018) to process the data ready for analysis was not completed. The reasons behind this change within the project have been justified in sections (Processing the Data) and (Discussion of Changes Made). The change was made to process the data using R (RStudio, 2018) instead as this would be able to handle the number of tweets collected. It also made sense to collect, store, cleanse, process and analyse the data all in one place. All other objectives were not changed and were completed successfully, outlined throughout the project documentation.

7.5 Summary of Chapter

From this chapter, it is clear to see the overall results from the analysis and areas which can be taken further for future work. Following this chapter, references are given as well as appendices which includes documents and code to help support the project.

8. Bibliography

Matta, Lunesu, Marchesi, n.d. *Bitcoin Spread Prediction Using Social and Web Search Media.* [pdf] Italy: University of Cagliari. Available at: <https://pdfs.semanticscholar.org/1345/a50edee28418900e2c1a4292ccc51138e1eb.pdf> [Accessed 02 April 2018].

Jermain Kaminski, 2014. *Nowcasting the Bitcoin Market with Twitter Signals.* [pdf] Cambridge, USA: MIT. Available at: <https://arxiv.org/pdf/1406.7577.pdf> [Accessed 02 April 2018].

Phillips, RP., Gorse, DG., 2017. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In: IEEE, *Computational Intelligence (SSCI).* Honolulu, USA, 2017.

Kim et al., 2016. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. *PLoS ONE* [e-journal] 11(8). https://doi.org/10.1371/journal.pone.0161197.

Satoshi, Nakamoto, 2009. *Bitcoin: A peer-to-peer Electronic Cash System.* [pdf] Bitcoin. Available at: < https://bitcoin.org/bitcoin.pdf> [Accessed 16 January 2018].

Kalampokis Evangelos, Tambouris Efthimios, Tarabanis Konstantinos, (2013) "Understanding the predictive power of social media", Internet Research, Vol. 23 Issue: 5, pp.544-559, https://doi.org/10.1108/IntR-06-2012-0114

Sabrina Bresciani and Andreas Schmeil, n.d. *Social Media Platforms for Social Good.* [pdf] IEEE. Available at: < https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6227944 > [Accessed 12 April 2018].

Sitaram Asur and Bernardo Huberman, 2010. *Predicting the Future With Social Media.* [pdf] IEEE. Available at: < https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5616710 > [Accessed 12 April 2018].

Hanna, R., Rohm, A., Crittenden, V., 2011. We're all connected: The power of the social media ecosystem. *Business Horizons*, [e-journal] 54 (3), pp. 265-273. https://doi.org/10.1016/j.bushor.2011.01.007

Ibrahim, A., et al., 2017. Analysis of Weakness of Data Validation from Social CRM. In: Sriwijaya University, *International Conference on Data and Software Engineering (ICoDSE)*. Palembang, Indonesia, 1-2 Nov 2017. IEEE.

Deng, X., 2017. Big Data Technology and Ethics Considerations in Customer Behavior and Customer Feedback Mining. In: Redmond, *IEEE International Conference on Big Data (Big Data).* Boston, USA, 11-14 Dec 2017. IEEE.

Deodhar, L., Divakaran, D., Gurusamy, M., 2017. Analysis of Privacy Leak on Twitter. In: National University of Singapore, *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. Singapore, 4-8 Dec 2017. IEEE.

Mahmood, S., 2012. New Privacy Threats for Facebook and Twitter Users. In: University College London, *2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Victoria, Canada, 12-14 Nov 2012. IEEE.

Trupthi, M., Pabboju, S., Narasimha, G., 2017. SENTIMENT ANALYSIS ON TWITTER USING STREAMING API. In: JNTUH and CBIT, *2017 IEEE 7th International Advance Computing Conference.* Hyderabad, India, 5-7 Jan 2017. IEEE

Kaur, H., Mangat, V., Nidhi., 2017. A Survey of Sentiment Analysis techniques. In: Panjab University, *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*. Palladam, India, 10-11 Feb 2017. IEEE.

Priyanka, D., Senthikumar, R., 2016. Sampling Techniques for Streaming Dataset using Sentiment Analysis. In: Anna University-MIT campus, *2016 FIFTH INTERNATIONAL CONFERENCE ON RECENT TRENDS IN INFORMATION TECHNOLOGY*. Chennai, India, 8-9 April 2016. IEEE.

Patti, V., Damiano, R., Bosco, C., 2017. Ethical Implications of Analyzing Opinions, Emotions and Interactions in Social Media. In: University Degli Studi di Torino, Italy, *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. San Antonio, USA, 23-26 Oct 2017. IEEE.

Joshi, R., Tekchandani, R., 2016. Comparative Analysis Of Twitter Data Using Supervised Classifiers. In: Thapar University, *2016 International Conference on Inventive Computation Technologies (ICICT).* Coimbatore, India, 26-27 Aug 2016. IEEE.

Nguyen, Q., 2016. Blockchain – A Financial Technology For Future Sustainable Development. In: University of Technical Education, *2016 3rd International Conference on Green Technology and Sustainable Development*. Kaohsiung, Taiwan, 24-25 Nov 2016. IEEE.

Bohr, J., Bashir, M., 2014. Who Uses Bitcoin? An exploration of the Bitcoin community. In: University of Illinois, *2014 Twelfth Annual Conference on Privacy, Security and Trust (PST).* Toronto, Canada, 23-24 July 2014. IEEE.

Vujičić, D., Jagodić, D., Ranđić, S., 2018. Blockchain Technology, Bitcoin, and Ethereum: A Brief Overview. In: University of Kragujevac, *17th International Symposium INFOTEH-JAHORINA*. East Sarajevo, Bosnia-Herzegovina, 21-23 March 2018. IEEE.

Lyudmyla, K., Vitalii, B., Tamara, R., 2017. Fractal Time Series Analysis of Social Network Activities. In: Kharkiv National University of Radioelectronics, *2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T).* Kharkov, Ukraine, 10-13 Oct 2017. IEEE.

Phillips, R., Gorse, D., 2017. Predicting Cryptocurrency Price Bubbles Using Social Media Data and Epidemic Modelling. In: University College London, *2017 IEEE Symposium Series on Computational Intelligence (SSCI).* Honolulu, USA, 27 Nov – 1 Dec 2017. IEEE.

He, S., et al., 2017. A Social-Network-Based Cryptocurrency Wallet-Management Scheme. *IEEE Access,* 6, pp. 7654 – 7663.

Radityo, A., Munajat, Q., Budi, I., 2017. Prediction of Bitcoin exchange rate to American dollar using artificial neural network methods. In: Universitas Indonesia, *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Bali, Indonesia, 28-29 Oct 2017. IEEE.

Saad, M., Mohaisen, A., 2018. Towards Characterizing Blockchain-based Cryptocurrencies for Highly-Accurate Predictions. In: University of Central Florida, *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* Honolulu, USA, 15-19 April 2018. IEEE.

Wang, Z., et al., 2016. Fine-Grained Sentiment Analysis of Social Media with Emotion Sensing. In: Institute of High Performance Computing (IHPC), *2016 Future Technologies Conference (FTC).* San Francisco, USA, 6-7 Dec 2016. IEEE.

Shahnawaz., Astya, P., 2017. Sentiment Analysis: Approaches and Open Issues. In: Saudi Electronic University, *2017 International Conference on Computing, Communication and Automation (ICCCA).* Greater Noida, India, 5-6 May 2017. IEEE.

Eyal, I., 2017. Blockchain Technology: Transforming Libertarian Cryptocurrency Dreams to Finance and Banking Realities. *Computer,* 50(9), pp. 38-49.

Mukhopadhyay, U., et al., 2016. A Brief Survey of Cryptocurrency Systems. In: Auburn University, *2016 14th Annual Conference on Privacy, Security and Trust (PST).* Auckland, New Zealand, 12-14 Dec 2016. IEEE.

Shehhi, A., Oudah, M., Aung, Z., 2014. Investigating Factors Behind Choosing a Cryptocurrency. In: Masdar Institute of Science and Technology, *2014 IEEE International Conference on Industrial Engineering and Engineering Management.* Bandar Sunway, Malaysia, 9-12 Dec 2014. IEEE.

Yuan, Y., Wang, F., 2018. Blockchain and Cryptocurrencies: Model, Techniques, and Applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9), pp. 1421-1428.

Nayak, A., Dutta, K., 2017. Blockchain: The Perfect Data Protection Tool. In: KIIT University, *2017 International Conference on Intelligent Computing and Control (I2C2)*. Coimbatore, India, 23-24 June 2017. IEEE.

Jain, A., Katkar, V., 2015. Sentiments analysis of Twitter data using data mining. In: Chinchwad College of Engineering, *2015 International Conference on Information Processing (ICIP).* Pune, India, 16-19 Dec 2015. IEEE.

Jose, R., Chooralil, V., 2016. Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach. In: Rajagiri School of Engineering and technology, *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. Ernakulam, India, 16-18 March 2016. IEEE.

Woldenmariam, Y., 2016. Sentiment analysis in a cross-media analysis framework. In: Science Umea University, *2016 IEEE International Conference on Big Data Analysis (ICBDA).* Hangzhou, China, 12-14 March 2016. IEEE.

Morstatter, F., Liu, H., Zeng, D., 2012. Opening Doors to Sharing Social Media Data. *IEEE Intelligent Systems*, 27(1), pp. 47-51.

Perikos, I., Haatzilygeroudis, I., 2018. A Framework for Analyzing Big Social Data and Modelling Emotions in Social Media. In: University of Patras, *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications.* Bamberg, Germany, 26-29 March 2018. IEEE.

Joseph, R., 2012. E-Government Meets Social Media: Realities and Risks. *IT Professional,* 14(6), pp. 9-15.

Mukkamala, R., et al, 2015. Detecting Corporate Social Media Crises on Facebook Using Social Set Analysis. In: Copenhagen Business School, *2015 IEEE International Congress on Big Data.* New York, USA, 27 June – 2 July 2015. IEEE.

Misra, G., Such, J., 2016. How Socially Aware Are Social Media Privacy Controls? *Computer* 49(3), pp. 96-99.

Baatarjav, E., Dantu, R., 2011. Current and Future Trends in Social Media. In: University of North Texas, *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing.* Boston, USA, 9-11 Oct 2011. IEEE.

Abirami, A., Gayathri, V., 2016. A survey on sentiment analysis methods and approach. In: Thiagarajar College of Engineering, Madurai, *2016 Eighth International Conference on Advanced Computing (ICoAC).* Chennai, India, 19-21 Jan 2017. IEEE.

Balaji, P., Nagaraju, O., Haritha, D., 2017. Levels of sentiment analysis and its challenges: A literature review. In: KL University, *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).* Chirala, India, 23-25 March 2017. IEEE.

Bhuiyan, H., et al., 2017. Retrieving YouTube video by sentiment analysis on user comment. In: University of Asia Pacific, *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA).* Kuching, Malaysia, 12-14 Sept 2017. IEEE.

Wang, Z., et al., 2014. Issues of Social Data Analytics with a New Method for Sentiment Analysis of Social Media Data. In: Institute of High Performance Computing, *2014 IEEE 6th International Conference on Cloud Computing Technology and Science.* Singapore, 15-18 Dec 2014. IEEE.

Save, A., Shekokar, N., 2017. Analysis of cross domain sentiment techniques. In: Sanghvi College of Engg, *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT).* Mysuru, India, 15-16 Dec 2017. IEEE.

Wang, Y., Rao, Y., Wu, L., 2017. A Review of Sentiment Semantic Analysis Technology and Progress. In: Xi'an Jiaotong University, *2017 13th International Conference on Computational Intelligence and Security (CIS).* Hong Kong, China, 15-18 Dec 2017. IEEE.

Wu, Y., et al., 2017. RIVA: A Real-Time Information Visualization and analysis platform for social media sentiment trend. In: National Taiwan University of Science and Technology, *2017 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT).* Munich, Germany, 6-8 Nov 2017. IEEE.

Shahare, F., 2017. Sentiment analysis for the news data based on the social media. In: Government College of Engineering, *2017 International Conference on Intelligent Computing and Control Systems (ICICCS).* Madurai, India, 15-16 June 2017. IEEE.

Chen, Y., Zhang, Z., 2018. Research on text sentiment analysis based on CNNs and SVM. In: Wuhan University of Science and Technology, *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. Wuhan, China, 31 May – 2 June 2018. IEEE.

Liang, P., Dai, B., 2013. Opinion Mining on Social Media Data. In: National Taiwan University of Science and

Technology, *2013 IEEE 14th International Conference on Mobile Data Management*. Milan, Italy, 3-6 June 2013. IEEE.

Chatterjee, R., Goyal, M., 2015. Tactics of Twitter Data Extraction for Opinion Mining. In: Manav Rachna College of Engineering, *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. New Delhi, India, 11-13 March 2015. IEEE.

Small, H., et al., 2012. What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data. *The International Journal of Digital Curation,* [e-journal] 7(1), pp. 174-197. http://dx.doi.org/10.2218/ijdc.v7i1.224.

Vela, S., Martinez, I., Reyes, L., 2012. Understanding risks, benefits, and strategic alternatives of social media applications in the public sector. *Government Information Quarterly,* [e-journal 29(4), pp. 504-511. https://doi.org/10.1016/j.giq.2012.07.002.

9. References

Bloom's Taxonomy, 1999. *Bloom's Taxonomy of Learning Domains.* [online] Available at: < http://www.nbna.org/files/Blooms%20Taxonomy%20of%20Learning.pdf > [Accessed 02 April 2018].

Bloom's Taxonomy, 1999. *Bloom's Taxonomy of Learning Domains.* [online image] Available at: < http://www.nwlink.com/~donclark/hrd/bloom.html > [Accessed 02 April 2018].

MindTools, n.d. *SMART Goals*. [online] Available at: < https://www.mindtools.com/pages/article/smart-goals.htm > [Accessed 02 April 2018].

Kaggle, 2018. *The home of data science & machine learning.* [online] Available at: < https://www.kaggle.com > [Accessed 04 April 2018].

RStudio, 2018. RStudio Desktop. (1.1.447). [computer program] R. Available at: < https://www.rstudio.com/products/rstudio/download/ > [Accessed 02 February 2018].

T. Graham., R. Ackland, 2017. SocialMediaLab. (0.23.2). [computer program] Available at: < https://cran.r-project.org/web/packages/SocialMediaLab/SocialMediaLab.pdf > [Accessed 04 April 2018].

Ethereum Project. 2018. *Ethereum Project*. [ONLINE] Available at: https://www.ethereum.org/. [Accessed 12 May 2018]

Ripple. 2018. *Ripple - One Frictionless Experience To Send Money Globally | Ripple*. [ONLINE] Available at: https://ripple.com/. [Accessed 12 May 2018].

Litecoin | Money for the Internet Age. 2018. *Litecoin | Money for the Internet Age*. [ONLINE] Available at: https://litecoin.com/. [Accessed 12 May 2018].

Statistical Analysis - What is it? | SAS UK. 2018. *Statistical Analysis - What is it? | SAS UK*. [ONLINE] Available at: https://www.sas.com/en_gb/insights/analytics/statistical-analysis.html. [Accessed 12 June 2018].

Facebook for Developers. 2018. *Facebook for Developers | To bring the world closer together.*. [ONLINE] Available at: https://developers.facebook.com/. [Accessed 12 June 2018].

LinkedIn. 2018. *Home | LinkedIn Developer Network*. [ONLINE] Available at: https://developer.linkedin.com/. [Accessed 8 June 2018].

Standard search API — Twitter Developers. 2018. *Standard search API — Twitter Developers*. [ONLINE] Available at: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html. [Accessed 10 June 2018].

YouTube, developers Google Developers. 2018. *YouTube Data API | Google Developers*. [ONLINE] Available at: https://developers.google.com/youtube/v3/. [Accessed 06 June 2018].

reddit: the front page of the internet. 2018. *reddit: the front page of the internet*. [ONLINE] Available at: https://www.reddit.com/. [Accessed 14 June 2018].

Tim Berners-Lee. 2018. *Tim Berners-Lee*. [ONLINE] Available at: https://www.w3.org/People/Berners-Lee/. [Accessed 20 July 2018].

AbdulMajedRaja, RS and Srivathsan K, 2017. Coinmarketcapr, (0.1). [computer program] Available at: < https://cran.r-project.org/web/packages/coinmarketcapr/coinmarketcapr.pdf > [Accessed 01 June 2018].

Jesse Vent, 2018. Crypto, (1.0.2). [computer program] Available at: < https://cran.r-project.org/web/packages/crypto/crypto.pdf > [Accessed 01 June 2018].

V. Spinu et al., 2018. Lubridate (1.7.4). [computer program] Available at: < https://cran.r-project.org/web/packages/lubridate/lubridate.pdf > [Accessed 01 June 2018].

B. Goodrich, D. Kurkiewicz and T. Rinker, 2018. Qdap, (2.3.0). [computer program] Available at: < https://cran.r-project.org/web/packages/qdap/qdap.pdf > [Accessed 01 June 2018].

M. Kearney, 2018. Rtweet, (0.6.7). [computer program] Available at: < https://cran.r-project.org/web/packages/rtweet/rtweet.pdff > [Accessed 30 June 2018].

M. Jockers, 2017. Syuzhet, (1.0.4). [computer program] Available at: < https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf > [Accessed 30 June 2018].

Sood, G., Lyons, K., Muschelli, J., 2018. Tuber, (0.9.7) [computer program] Available at: < https://cran.r-project.org/web/packages/tuber/tuber.pdf > [Accessed 30 June 2018].

Michael Piccirilli, 2016. Rlinkedin. (0.2). [computer program] Available at: < https://cran.r-project.org/web/packages/Rlinkedin/Rlinkedin.pdf > [Accessed 04 April 2018].

Ying Chen, 2016. *Convert .json to .csv*. [online] Available at: < http://www.yingchen.live/convert-json-csv/ > [Accessed 16 April 2018].

MongoDB, 2018. *Mongo DB.* [online] Available at: < https://www.mongodb.com/ > [Accessed 13 April 2018].

MongoDB, 2018. *Spark Connector R Guide*. [online] Available at: < https://docs.mongodb.com/spark-connector/master/r-api/ > [Accessed 16 April 2018].

Hortonworks, 2018. Sandbox. [computer program] Hortonworks. Available at: < https://hortonworks.com/products/sandbox/ > [Accessed 02 April 2018].

Apache Spark, 2018. Spark. (2.3.0). [computer program] Apache. Available at: < https://spark.apache.org/ > [Accessed 13 January 2018].

Brandwatch, 2015. *Understanding Sentiment Analysis: What It Is & Why It's Used.* [online] Available at: < https://www.brandwatch.com/blog/understanding-sentiment-analysis/ > [Accessed 18 April 2018].

Tableau, 2016. *Tableau Desktop*. (10.0). [computer program] Tableau. Available at:< http://www.tableau.com/products/desktop> [Accessed 03 December 2016].

CoinAPI, 2018. *API.* [online] Available at: < https://www.coinapi.io/# > [Accessed 15 April 2018].

CoinMarketCap, 2018. *API*. [online] Available at: < https://coinmarketcap.com/api/ > [Accessed 15 April 2018].

Brandwatch, 2015. *Brandwatch for students.* [online] Available at: < https://www.brandwatch.com/students/ > [Accessed 18 April 2018].

British Computer Society, 1957. *History of BCS*. [online] Available at: <http://www.bcs.org/category/11284> [Accessed 10 December 2016].

British Computer Society, 1957. *BCS Code of Conduct*. [pdf] BCS. Available at: < http://www.bcs.org/upload/pdf/conduct.pdf> [Accessed 10 December 2016].

Parliament, 1991. *Acts.* [online] Available at: <http://www.parliament.uk/about/how/laws/acts/> [Accessed 10 December 2016].

Data Protection Act 1998. (c.29). London: HMSO.

EU General Data Protection Regulation. *General Data Protection Regulation (GDPR).*

Computer Misuse Act 1990. (c.18). London: HMSO.

Intellectual Property Act 2014. (c.18). London: HMSO.

Copyright Act 1956. (c.74). London: HMSO.

Gabour Terstyanszky, GT., 2017. *Big_data_analytics*. [image] Available at: Westminster University Blackboard [Accessed 02 February 2018].

Gabour Terstyanszky, GT., 2017. *Big_data_mapreduce_02*. [image] Available at: Westminster University Blackboard [Accessed 02 February 2018].

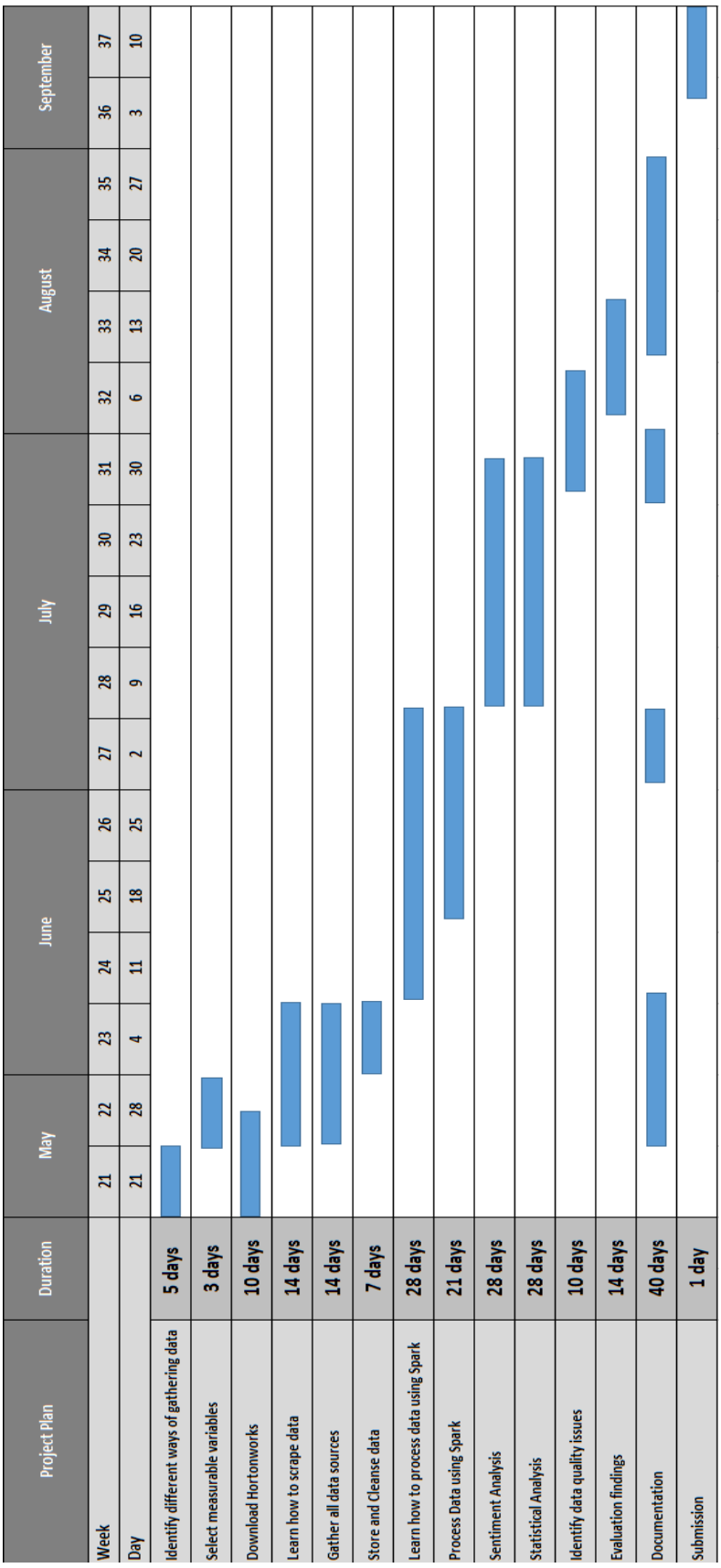
Forbes, 2018. *Blockchain adoption: How Close Are We Really?* [image] Available at: < https://www.forbes.com/sites/luuloi/2018/01/26/blockchain-adoption-how-close-are-we-really/#1f989648d9dc> [Accessed 02 February 2018].

Japan Times, 2018. *Japan Global Leader Cryptocurrency Investment.* [image] Available at: <https://www.japantimes.co.jp/news/2018/01/23/business/japan-global-leader-cryptocurrency-investment/#.WqxAMWrFKM8> [Accessed 02 February 2018].

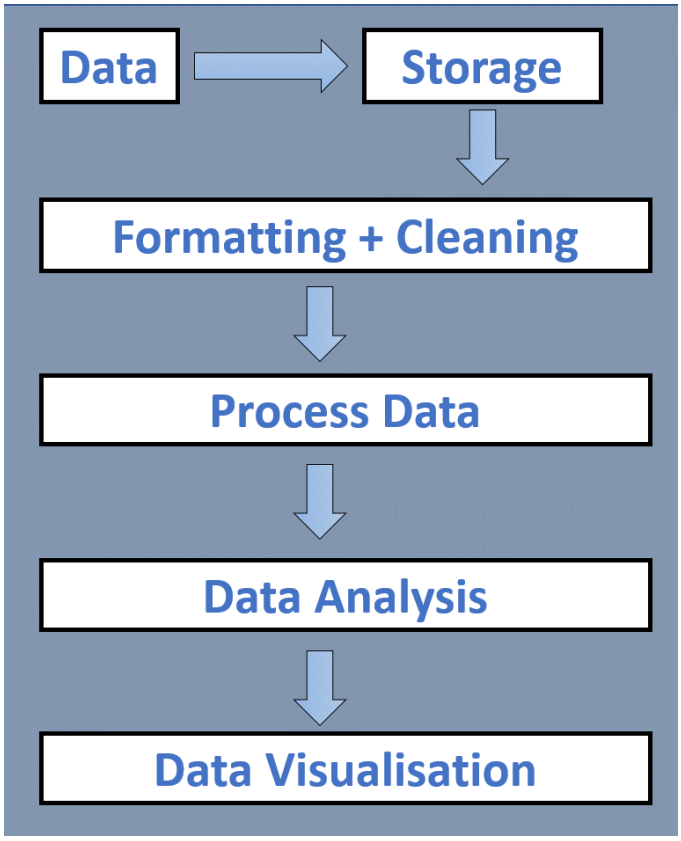
CryptoDataDownload, 2018. *Kraken Market.* [online] Available at: < http://www.cryptodatadownload.com > [Accessed 02 June 2018].

10. Appendices

Appendix A – Project Schedule



Appendix B – Project Methodology



Appendix C – Analysis and Visualisation Code

This part of the appendix is just a snippet from the R code of an example of just one of the coins. The example shows data analysis and visualisations and creating csv for Litecoin Twitter data. The same steps were copied for each coin and for the YouTube data.

“

```
#################################################################
##############___SENTIMENT + POLARITY STEPS___#############
#################################################################
### installs the package
install.packages("syuzhet")
library(syuzhet)

#gets polarity + sentiment scores and stores them in a new data frame
LITECOIN_PSscores = get_nrc_sentiment(LITECOIN_2)

#divides polarity and sentiment
LITECOIN_POLARITY = LITECOIN_PSscores[ ,9:10]
LITECOIN_SENTIMENT = LITECOIN_PSscores[ ,1:8]

###LITECOIN POLARITY
### calculates the total sum of each polarity storing it in a new data frame
sum(LITECOIN_POLARITY$negative)
sum(LITECOIN_POLARITY$positive)
SUM_POL_LITECOIN = data.frame(sum(LITECOIN_POLARITY$negative),
sum(LITECOIN_POLARITY$positive))
SUM_POL_LITECOIN[0, ]
names(SUM_POL_LITECOIN) [1] <- "Negative"
names(SUM_POL_LITECOIN) [2] <- "Positive"
SUM_POL_LITECOIN[0, ]

###LITECOIN SENTIMENT
###creates a dataframe was the total scores of each emotion
SUM_SENT_LITECOIN = data.frame(sum(LITECOIN_SENTIMENT$anger),
                    sum(LITECOIN_SENTIMENT$anticipation),
                    sum(LITECOIN_SENTIMENT$disgust),
                    sum(LITECOIN_SENTIMENT$fear),
                    sum(LITECOIN_SENTIMENT$joy),
                    sum(LITECOIN_SENTIMENT$sadness),
                    sum(LITECOIN_SENTIMENT$surprise),
                    sum(LITECOIN_SENTIMENT$trust))
```

```
names(SUM_SENT_LITECOIN) [1]  <- "Anger"
names(SUM_SENT_LITECOIN) [2]  <- "Anticipation"
names(SUM_SENT_LITECOIN) [3]  <- "Disgust"
names(SUM_SENT_LITECOIN) [4]  <- "Fear"
names(SUM_SENT_LITECOIN) [5]  <- "Joy"
names(SUM_SENT_LITECOIN) [6]  <- "Sadness"
names(SUM_SENT_LITECOIN) [7]  <- "Surprise"
names(SUM_SENT_LITECOIN) [8]  <- "Trust"

SUM_SENT_LITECOIN[0, ]
SUM_SENT_LITECOIN

#################################################################
######################___VISULISATIONS___#####################
#################################################################
###installs the package
install.packages("ggpplot2")
library(ggplot2)
#CHANGES INTO A MATRIX TO BE ABLE TO DISPLAY THE DATA
SUM_POL_LITECOIN = data.matrix(SUM_POL_LITECOIN, rownames.force = TRUE)
barplot(SUM_POL_LITECOIN)

SUM_SENT_LITECOIN = data.matrix(SUM_SENT_LITECOIN, rownames.force =
TRUE)
barplot(SUM_SENT_LITECOIN)

boxplot(LITECOIN_SENT)
summary(LITECOIN_SENT)

ggplot(data = LITECOIN_PRICES, aes(x=Date, y=Close)) + geom_step(colour = "red",
size = 0.6)

#########################
####___LINE GRAPHS___####
#########################

#simple
ggplot(data = LITECOIN_PRICES, aes(x=Date, y=Close)) + geom_line()
```

```
#PLOTS SIMPLE LINE GRAPH OF DATA
ggplot(data = LITECOIN_PRICES, aes(x=Date, y=Close)) +
  geom_line(colour = "white", size = 0.6) +
  labs(title = "Litecoin Prices", y="Closing Prices") +
  theme_dark() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

###LITECOIN
ts_plot(LITECOIN, by = "mins")
ts_plot(LITECOIN, by = "hours")
ts_plot(LITECOIN, by = "days")

################################
####___MULTIPLE DISPLAYS___####
################################

nstall.packages("gridExtra")
library(gridExtra)

p3 = ts_plot(LITECOIN, by = "hours")
p2 = ggplot(data = LITECOIN_PRICES, aes(x=Date, y=Close)) +
  geom_line(colour = "white", size = 0.6) +
  labs(title = "Litecoin Prices", y="Closing Prices")
p4 = ts_plot(LITECOIN_YOUTUBE, by = "hours")

### This plots the number of tweets one graph
###the second graph underneath is the prices
###the third graph underneath is number of youtube videos
grid.arrange(p3, p2, p4, nrow = 3)

#################################################################
#########___GROUP ELEMENT SCORES BY EVERY HOUR___#############
#################################################################

###The code below groups creates a new file for every emotion
#in that new file it groups the scores together by the hour
#so then this is easy to analyse against prices as the price
#of cryptocurrencies is one value every hour
```

```
Lsum_Pos = aggregate(LITECOIN00['positive'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Neg = aggregate(LITECOIN00['negative'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Ang = aggregate(LITECOIN00['anger'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Ant = aggregate(LITECOIN00['anticipation'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Dis = aggregate(LITECOIN00['disgust'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Fea = aggregate(LITECOIN00['fear'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Joy = aggregate(LITECOIN00['joy'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Sad = aggregate(LITECOIN00['sadness'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Sur = aggregate(LITECOIN00['surprise'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)
Lsum_Tru = aggregate(LITECOIN00['trust'], list(cut(LITECOIN00$created_at, "1 hour")), sum, na.rm=TRUE)

LITECOIN_1HR_SCORE = data.frame(Lsum_Pos$positive,
                        Lsum_Neg$negative,
                        Lsum_Ang$anger,
                        Lsum_Ant$anticipation,
                        Lsum_Dis$disgust,
                        Lsum_Fea$fear,
                        Lsum_Joy$joy,
                        Lsum_Sad$sadness,
                        Lsum_Sur$surprise,
                        Lsum_Tru$trust,
                        TIMEFRAME$Date)

write.csv(LITECOIN_1HR_SCORE, file="Litecoin_1hr_Scores.csv")
write.csv(LITECOIN_PRICES, file="LITECOIN_PRICES.csv")

###creating number of tweets per crypto
NUMBER_OF_TWEETS = data.frame(nrow(LITECOIN_2),
                        nrow(RIPPLE),
                        nrow(ETHEREUM),
```

```
                    nrow(BITCOIN))
NUMBER_OF_TWEETS
names(NUMBER_OF_TWEETS) [1]  <- "Litecoin"
names(NUMBER_OF_TWEETS) [2]  <- "Ripple"
names(NUMBER_OF_TWEETS) [3]  <- "Ethereum"
names(NUMBER_OF_TWEETS) [4]  <- "Bitcoin"

NUMBER_OF_TWEETS = data.matrix(NUMBER_OF_TWEETS, rownames.force = TRUE)
barplot(NUMBER_OF_TWEETS)”
```