Institiúid Teicneolaíochta Cheatharlach

INSTITUTE of TECHNOLOGY CARLOW

At the Heart of South Leinster

Project Report

Prediction of cryptocurrency price movement using sentiment analysis and machine learning

By Josh Browne
College ID: C00224660
Date: 2021

Project Abstract

In this study, the predictability of cryptocurrencies prices are analyzed at the daily and hourly level, through the use of machine learning methods including the use of sentiment analysis and LSTM Recurrent Neural Networks. Cryptocurrencies are becoming increasingly relevant in the financial world and can be considered as an emerging market. The high volume of data availability of the cryptocurrency market makes it an enticing area of study, from which it is possible to derive insights into the behaviour of the markets through analyzing trends and impacting factors, such as the public’s online sentiment. Essentially, this project carries out an investigation to assess the relationship between cryptocurrency closing price trends and online sentiment trends. The calculations for the online sentiment is done by retrieving tweets from Twitter and performing sentiment analysis on the textual data. There is an implementation to train a Recurrent Neural Network to make predictions for a future closing price of a cryptocurrency, using elements of the price and sentiment data to train the model. The predicted results and the calculated sentiment data along with some other useful data is displayed on this project's web application, which was the technical parts to the work done.

Project Introduction:

The popularity of cryptocurrencies had grown greatly in 2017 due to several consecutive months of exponential growth of their market capitalization [1], it has continued growth on the large scale and peaked at more than $1.8 trillion in 2021. Today, there are more than 1,900 actively traded cryptocurrencies. There are estimated between 2.9 and 5.8 million private as well as institutional investors in the different transaction networks, according to a recent survey [2], and access to the market has become easier over time with the growing number of online exchanges. Today, cryptocurrencies can be bought using fiat currency from a number of online exchanges (e.g Binance [3], Coinbase [4], Kraken [5] etc.), as the easability of purchasing cryptocurrency increases through these mediums it seems the market's popularity grows. The daily trading volume for many of the top exchanges is in the billions, at the time of writing, Binance’s daily trading volumes are around $70 billion. Recent reports have shown that there are around 150 active cryptocurrency hedge funds, which collectively hold around $1 billion assets under management, this excludes crypto index funds and crypto venture funds [6].

The market is diverse and provides investors with a variety of different products. To mention a few, Bitcoin was expressly designed as a medium of exchange [7]. Ethereum is a public, blockchain-based distributed computing platform featuring smart contract (scripting) functionality, and Ether is a cryptocurrency whose blockchain is generated by the Ethereum platform [8]. Ripple is a real-time gross settlement system (RTGS), currency exchange, and remittance network Ripple [9], and IOTA is focused on providing secure communications and payments between agents on the Internet of Things [10].

The emergence of a self-organised market of virtual currencies and/or assets whose value is generated primarily by social consensus [11] has naturally attracted interest from the scientific community [12]. There have been various studies that have focused on the analysis and forecasting of price fluctuations, using mostly traditional approaches for financial markets analysis and prediction [13].

The success of machine learning techniques for stock markets prediction [14] suggests that these methods could be effective also in predicting cryptocurrencies prices. Previous attempts at applying machine learning techniques on time series data include using random forests[15], Bayesian neural network[16], Support Vector Machine (SVM) [17] and many more methods have been attempted to varying success. These studies were able to anticipate, to different degrees, the price fluctuations of Bitcoin, and revealed that best results were achieved by neural network based algorithms.

Twitter is a popular microblogging service where users create status messages, called "tweets". These tweets sometimes express opinions about different topics [18]. Sentiment analysis on Tweets is a field that has recently attracted research interest. Sentiment analysis in Twitter tackles the problem of analyzing the tweets in terms of the opinion they express [19].

In this project, sentiment analysis along with machine learning principles are applied to find the correlation between “public sentiment” and “market sentiment”. Twitter data is used to calculate a numerical value representing the public mood, this mood value is used along with previous days’ mood values to predict future market movements.

Literature Review:

A Deeper Dive Into Cryptocurrency And Its Goals:

Amongst the many controversies surrounding cryptocurrencies, a popular topic of debate is whether it should be classified as a commodity, investment, property, currency or digital currency. Bitcoin puts cryptocurrencies center stage in the popular press [20]. Bitcoin managed to grow 14x in 2017, starting the year at $998.33 and had reached $14,156.40 by the beginning of 2018. A brief history of Bitcoin’s extreme volatility can be viewed in Fig. 1. Bitcoin is the first successful cryptocurrency, it was created in January 2009 and it utilizes a technology called blockchain, which is a combination of cryptography, consensus algorithms, economic incentives and a publicly distributed ledger to secure its transactions.

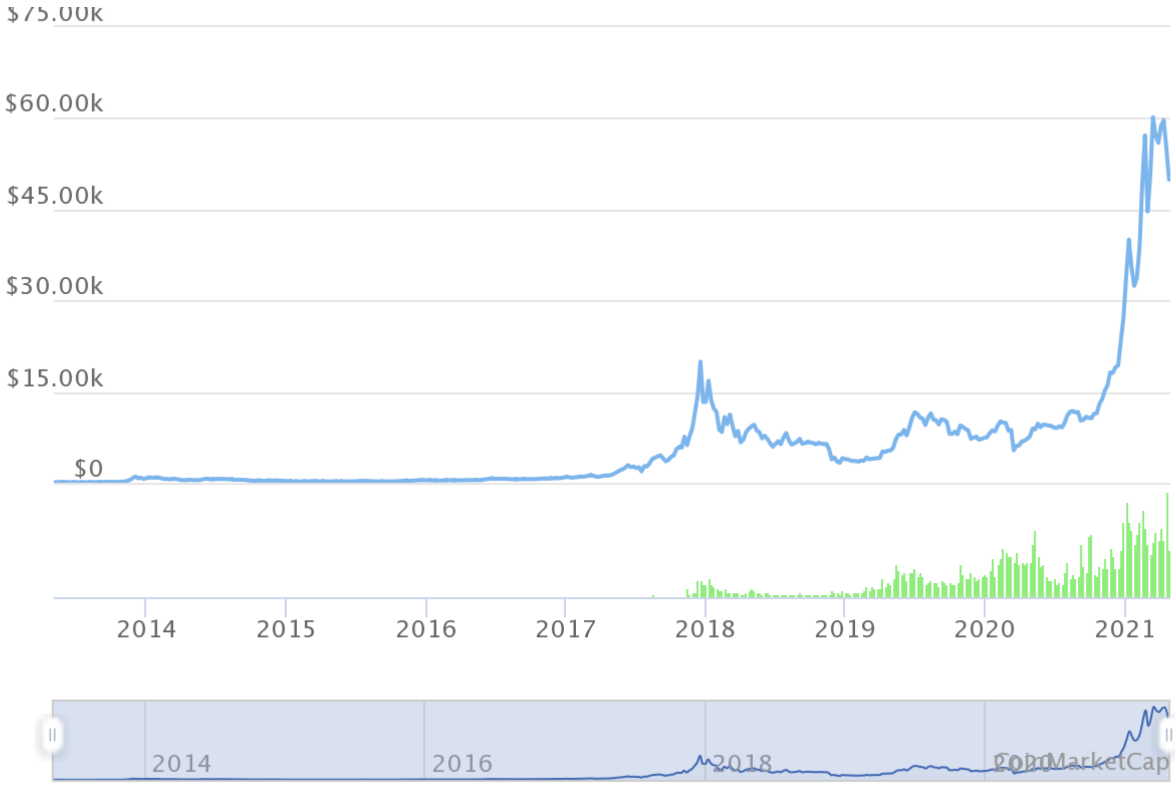
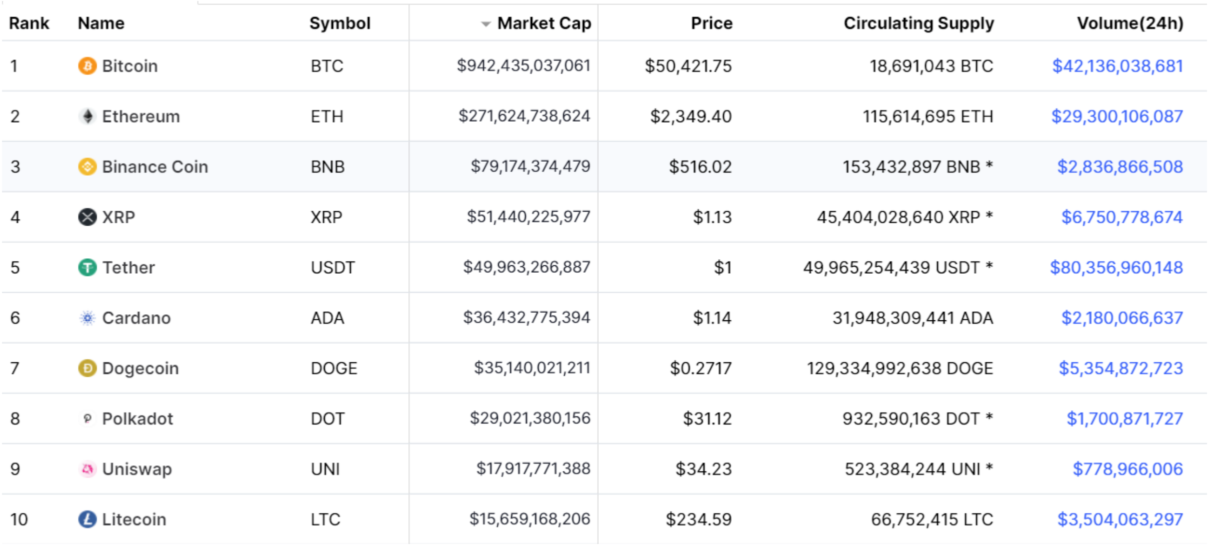


Figure 1: Bitcoin price from 2013 to April 2021

There have been many cryptocurrencies created after Bitcoin, but Bitcoin continues to be the most popular as it retains the top of the list with the largest market capitalization and trading volume, as can be viewed in Table 1.

Ethereum is the longtime standing second most popular cryptocurrency on the market. Ethereum is a project which attempts to build the generalised technology; technology on which all transaction based state machine concepts may be built. Moreover it aims to provide to the end-developer a tightly integrated end-to-end system for building software on a hitherto unexplored compute paradigm in the mainstream: a trustful object messaging compute framework [21]. One of the many goals of this project is to facilitate transactions between consenting individuals who would otherwise have no means to trust one another.



Rank	Name	Symbol	Market Cap	Price	Circulating Supply	Volume(24h)
1	Bitcoin	BTC	$942,435,037,061	$50,421.75	18,691,043 BTC	$42,136,038,681
2	Ethereum	ETH	$271,624,738,624	$2,349.40	115,614,695 ETH	$29,300,106,087
3	Binance Coin	BNB	$79,174,374,479	$516.02	153,432,897 BNB *	$2,836,866,508
4	XRP	XRP	$51,440,225,977	$1.13	45,404,028,640 XRP *	$6,750,778,674
5	Tether	USDT	$49,963,266,887	$1	49,965,254,439 USDT *	$80,356,960,148
6	Cardano	ADA	$36,432,775,394	$1.14	31,948,309,441 ADA	$2,180,066,637
7	Dogecoin	DOGE	$35,140,021,211	$0.2717	129,334,992,638 DOGE	$5,354,872,723
8	Polkadot	DOT	$29,021,380,156	$31.12	932,590,163 DOT *	$1,700,871,727
9	Uniswap	UNI	$17,917,771,388	$34.23	523,384,244 UNI *	$778,966,006
10	Litecoin	LTC	$15,659,168,206	$234.59	66,752,415 LTC	$3,504,063,297

Table 1: CryptoCurrency Market Capitalizations
Source: https://coinmarketcap.com/

The Benefits And Use Cases Of Sentiment Analysis:

The study of public opinion can provide us with valuable information and the various methods of analyzing sentiment has a wide range of applications. Sentiment Analysis is a type of data mining where you measure the inclination of people's opinions by using NLP (natural language processing), text analysis and computational linguistics. Computational Linguistics involves looking at the ways in which a machine would treat natural language. In other words, dealing with or constructing models for language that can allow for goals such as accurate machine translation of language, or the simulation of artificial intelligence. Sentiment Analysis is to use this technology (and others alike) to determine and dig subjective information from source materials.

Twitter is a great place for performing sentiment analysis because you can get public opinion on any topic through this platform. Twitter is the most popular microblogging platform for the task of Sentiment Analysis.

"Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Therefore

microblogging web-sites are rich sources of data for opinion mining and sentiment analysis.”
- Twitter as a Corpus for Sentiment Analysis and Opinion Mining [22]

The Sentiment Analysis process aims to determine how a certain person or group reacts to a given topic. They react because they are either interested or involved. And, these reactions go to none other than their social media accounts which makes social media as one of the leading platforms on the internet where anyone can do sentiment analysis.

Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large. Twitter’s audience varies from regular users to celebrities, company representatives, politicians and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.

Previous Implementations Of Sentiment Analysis:

Sentiment Analysis is studied at many different levels. In [23], authors describe their implementation of sentiment analysis for the purpose of gaining insight to the public’s opinion about *Tesla’s* product, the *CyberTruck*. This product is a polarizing electric pickup truck that has made itself known for its peculiar shape and its controversial unveiling event. Their efforts gave them the ability to learn about some key factors about how people generally viewed the product. Through analyzing the found data, they were able to conclude that this unveiling event and other events alike, contributed to make this model really popular and largely talked about. The team used multiple methods of textual analysis such as:

1. Monthly analysis of single words, assessing their frequency and sentiment value through a lexicon-emotion association.
2. Identification of relevant topics of discussion in tweets through the use of a probabilistic method called Latent Dirichlet Allocation or LDA.
3. A special dictionary was used that could give value to slangs, emoticons, emoji, valence shifters, amplifiers, adversatives and question weights, thus capturing more information contained in a sentence and deepening the level of analysis. [23]

For all of the above methods, visualization tools were implemented to better highlight the findings of these analyses. One key benefit to this work would be that a company could gain valuable knowledge on where best to focus their social media marketing based on which social media’s house the most interest for their products.

In a very recent article [24], the authors carried out research into the application of sentiment analysis for the purpose of fighting the highly infectious disease COVID-19. At the time of this research was carried out, news about the virus was spreading all over social media websites. Wherefore, these social media outlets were experiencing and presenting different views, opinions and emotions during various outbreak incidents. The focus point for this team's analysis was that of articles written about the newly recognised disease, along with articles written in the past about other infectious diseases. There were comparisons made to the public reaction of COVID-19 against previous pandemics caused by infectious disease. The methods used in this research were lexicon-based models, machine learning based models and hybrid based models. Interesting patterns were observed in the literature and the identified articles were grouped accordingly. Similar studies relating to the health sector are scarce, as detailed in the below excerpt.

"*To the best of our knowledge, previous related studies have focused on one kind of infectious disease. No previous study has examined multiple diseases via sentiment analysis.*" - [24]

There are large amounts of data found online and in published articles, this data can be a valuable asset for understanding people's sentiments regarding current events. With keeping this in mind, it is therefore not too hard to imagine a future where the technologies used for calculating sentiment increase in accuracy and these types of application methods become more relevant.

Previous Work Related To Finance:

For the topic of using machine learning as a means to make predictions on future stock values, there are a wide range of methodologies which have been used in the past and detailed in research papers. Before embarking on my work with this project, it was important for me to carry out research into these various methods in an effort to gain understanding of which route to take my project so that I would have the best chance at success with my machine learning model. For the remainder of this section I will be detailing some of my findings and how they impacted my decision making for this project.

There have been far more efforts towards forecasting the price of the stock market, rather than forecasting the cryptocurrency market, but since the two markets have similar impacting factors, much of my research was carried out on the more successful workings in forecasting the stock market. The key similarities between these markets are:

- The values of both markets are usually in fiat currencies.
- The level of demand and supply determines the price of stocks and Cryptocurrency. The higher the number of people requesting a stock or Cryptocurrency, the higher their price and value.
- They both involve high risk, the price could fall drastically and result in an investor losing large sums of money.

Due to these similarities, it can be surmised that some of these successful implementations focused on training machines to find trends in stock price fluctuations, might also provide similar success once adapted to the cryptocurrency market. One of the notable differences between these markets is that the crypto market has historically been far more volatile, meaning the prices are liable to change rapidly.

When comparing the various methods used for stock and crypto market predictions, it is clear that neural networks have the highest accuracy. The utilization of neural networks in exchange forecasting problems is extremely provoking because of some of their exceptional qualities. Firstly, neural networks show a striking capacity to extract context from convoluted or estimated information. They can be used to derive patterns and identify trends that are far too convoluted to be comprehended by humans or other conventional computer processes. Secondly, neural networks exhibit a nonlinear nature and are favored over the conventional linear models. Thirdly, a neural network that is trained on a specific dataset of a specific domain, can be easily re-trained to another dataset to forecast at similar levels of conditions [25]. That last point is particularly important to this project as the end goal is to build numerous models focused on many different cryptocurrencies, for the purpose of comparison as I expect the different currency communities to have different levels of online presence, resulting in different measurements of accuracy. Also, when the framework in review is continuously changing and updating, neural networks have the capability to accordingly alter the weights, hence adapt to the changes. The cryptocurrency market is a highly non-linear, dynamic and perpetually evolving framework. The aforementioned features of neural networks makes them the ideal answer for forecasting price's of the cryptocurrency market.

In recent times, the research world has seen many different algorithms and different types of neural networks being created. Thus, finding the ideal neural network for one's research has become a difficult task. It requires large amounts of research and analysis. In a recent thesis [26], comparisons were made for the predicting power of Support Vector Regression, Feed Forward Neural Networks and Convolutional Neural Networks which all had comparable results with some pros and cons of each. The Echo State Network was also found to have noisy behaviour in this thesis.

There has been some proven success for time series prediction through the use of Artificial Recurrent Neural Networks(RNN) and Long Short-Term Memory(LSTM). Hengjian Jia found that LSTMs learn patterns effective for stock market prediction and he obtained decent root-mean-square errors with different architectures of LSTM [27]. These studies helped me gain direction as to what technologies would be best suited for the use case for this project.

From studying the papers mentioned above, I chose to take my project down the route of using an LSTM model as it seemed to yield the best results for the problem of forecasting volatile market price. It was not my plan to simply implement a tried and tested method, but to also add in a less popular feature to the training dataset. This feature was that of public sentiment from Twitter users. A large point of interest in the cryptocurrency market is the large scale of available public sentiment data, particularly from social networks. This data can presumably be used to infer future human behaviour, and therefore could be used to develop advantageous trading strategies [28], as has been shown in recent attempts to detect speculative bubbles in the cryptocurrency market using sentiment analysis [29]. As of now, there have been few studies carried out which attempt to create profitable trading strategies in the cryptocurrency market, far less than that of stock market strategies at least. These findings encouraged me to explore this field in an attempt to further the research in this new and exciting market.

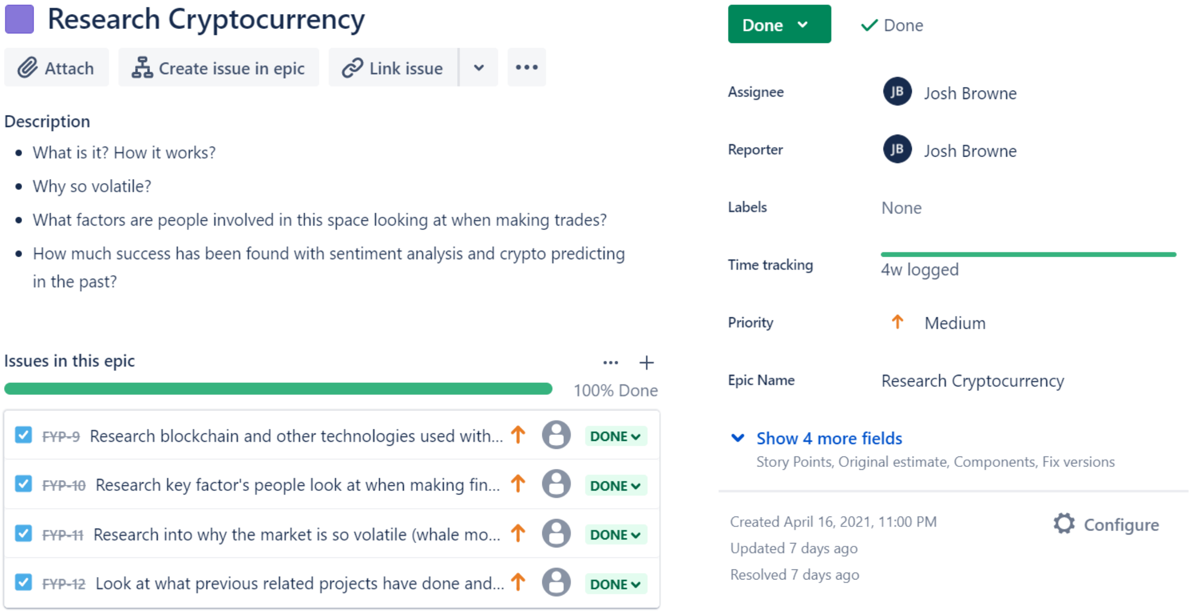
Evaluation and Discussion

The results in my findings with the LSTM model can be found on my website. The models predictions do seem to have some sort of predictive power, especially when it comes to the big rises and falls in prices. It is difficult to find the true benefits of using my sentiment data as a training feature without directly comparing it with a model trained without sentiment data as an input feature. It was due to time constraints that I could not make these comparisons

Project Milestones

I feel that overall the planning and work execution for this project went well. There were some issues toward the final days with building my models and fixing this issue was far more time consuming than anticipated. I used Jira to plan out my features and tasks, but as I cannot change the access permissions with a free account, I will detail my schedule below and include screenshots of the Jira features/tasks.

September:
During this month, the bulk of project related work done was to carry out research and try to nail down the core concepts to my project. At this point I had decided on implementing sentiment analysis on something practical, and had an idea of possibly implementing it for use in the cryptocurrency space. I had very limited knowledge about cryptocurrency at this point, so I carried out research of cryptocurrency in general to get a better overview of the space and I also looked into possible ways of using sentiment analysis for use in machine learning models during this time.



Research Cryptocurrency

Description

- What is it? How it works?
- Why so volatile?
- What factors are people involved in this space looking at when making trades?
- How much success has been found with sentiment analysis and crypto predicting in the past?

Issues in this epic

100% Done

☑ ~~FYP-9~~ Research blockchain and other technologies used with... ↑ DONE

☑ ~~FYP-10~~ Research key factor's people look at when making fin... ↑ DONE

☑ ~~FYP-11~~ Research into why the market is so volatile (whale mo... ↑ DONE

☑ ~~FYP-12~~ Look at what previous related projects have done and... ↑ DONE

✓ Done

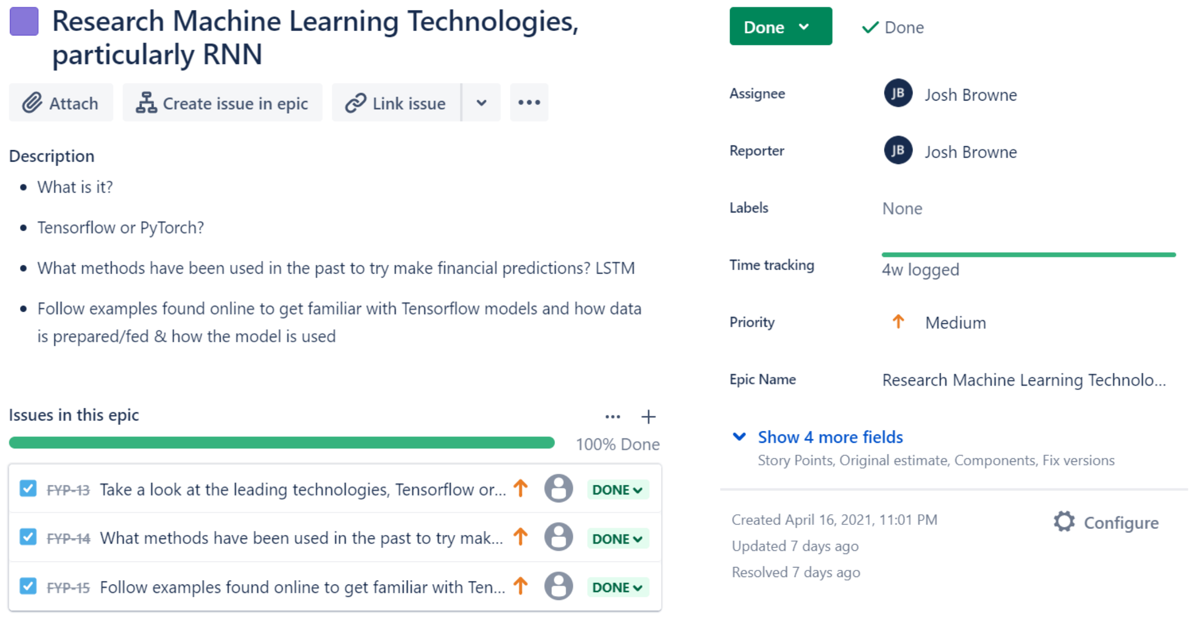
Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	4w logged
Priority	↑ Medium
Epic Name	Research Cryptocurrency

Story Points, Original estimate, Components, Fix versions

Created April 16, 2021, 11:00 PM
Updated 7 days ago
Resolved 7 days ago

October:

Now that I had a rough idea of what I wanted my project to be, it was time to dive into some research for machine learning techniques. During this time I looked into both Tensorflow and PyTorch technologies, eventually discovering that Tensorflow was the more suitable option, partly because it is more popular and so arguably more documentation and help can be found online pertaining to it. I spent time researching what methods have been used in the past to make time series predictions, and quickly found that neural networks and particularly Long Short Term Memory(LSTM) models seemed the most promising for this task. I then followed a few basic tutorials where I could familiarize myself with the Tensorflow framework.



Research Machine Learning Technologies, particularly RNN

Description

- What is it?
- Tensorflow or PyTorch?
- What methods have been used in the past to try make financial predictions? LSTM
- Follow examples found online to get familiar with Tensorflow models and how data is prepared/fed & how the model is used

Issues in this epic

100% Done

☑	~~FYP-13~~	Take a look at the leading technologies, Tensorflow or...	DONE
☑	~~FYP-14~~	What methods have been used in the past to try mak...	DONE
☑	~~FYP-15~~	Follow examples found online to get familiar with Ten...	DONE

✓ Done

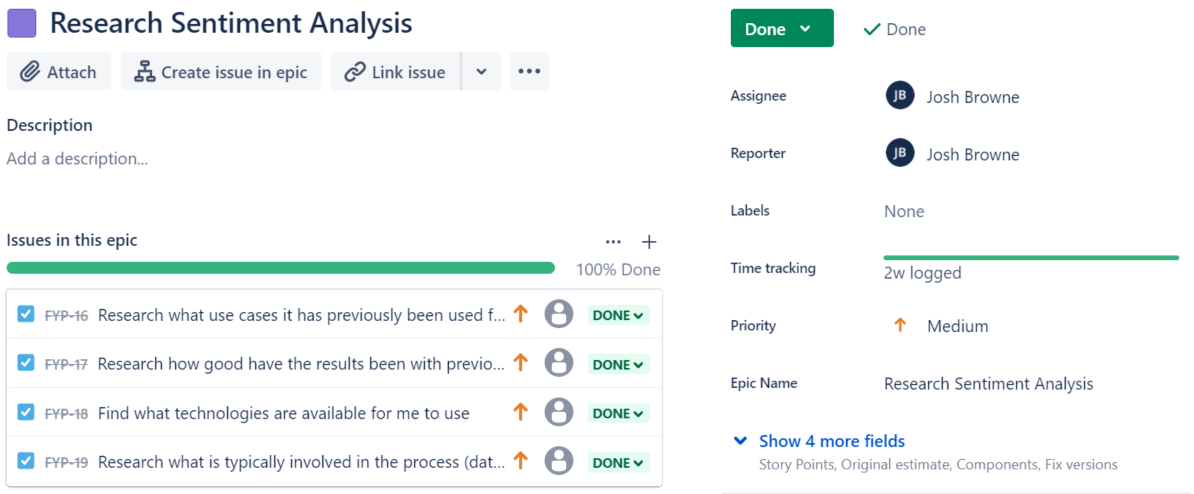
Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	4w logged
Priority	Medium
Epic Name	Research Machine Learning Technolo...

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

Created April 16, 2021, 11:01 PM
Updated 7 days ago
Resolved 7 days ago

November:

During this month I focused on carrying out research on tools to use for Sentiment Analysis. I found there was a wide selection of open source tools that can be utilised to perform sentiment analysis, but decided that the *TextBlob* and *NLTK* technologies were most suitable as they were more popularly used and sources I found described them as having a wider lexicon library. I also spent some time investigating IBM Watson's [30] offering of sentiment analysis tools, primarily because they boast a new and interesting type of sentiment analysis which is context specific. In order to use this context specific sentiment analysis, you need a dataset of popular sentences and phrases which are marked either positive or negative respectfully, that relate to the topic you plan on using the tool for. I found this intriguing but could not find a suitable dataset anywhere online. I did consider creating a dataset myself but as I would need to mark each sentence/phrase positive or negative by hand, my supervisor advised me to not pursue this route as it would not be scalable and would be far too time consuming. In the end I focused on researching what was involved with using TextBlob and NLTK tools and implemented sentiment scoring for some junk textual data to get familiar with these tools.



Research Sentiment Analysis

Description

Add a description...

Issues in this epic

100% Done

☑	~~FYP-16~~	Research what use cases it has previously been used f...	↑	DONE
☑	~~FYP-17~~	Research how good have the results been with previo...	↑	DONE
☑	~~FYP-18~~	Find what technologies are available for me to use	↑	DONE
☑	~~FYP-19~~	Research what is typically involved in the process (dat...	↑	DONE

Done ✓ Done

Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	2w logged
Priority	↑ Medium
Epic Name	Research Sentiment Analysis

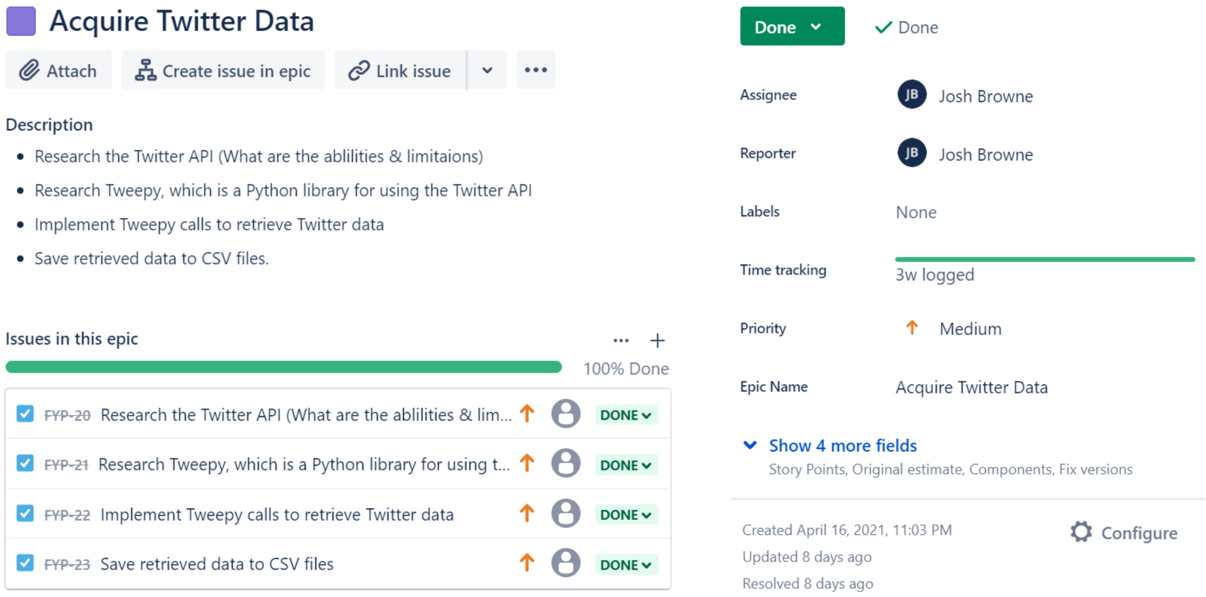
Show 4 more fields

Story Points, Original estimate, Components, Fix versions

December/January

During this stage of the project development, I was focusing on retrieving Twitter data. To do this I needed to request a developer account from Twitter, they have a free version that they allow researchers and students access to. Once I got my developer account access keys I was able to make requests to the Twitter API through the Python API for Twitter called Tweepy. During this stage of the development, I made requests to the Twitter database and retrieved data successfully. I spent time familiarising myself with the tool's capabilities and limitations. I found there were limitations on how many data points could be retrieved but that this would not pose much of an issue as the request limitations reset every 15 minutes and I could acquire the data I need over a period of time. The limitation with this that would prove to be an issue was that I could only retrieve data from the past 30 days. This meant that in the end, my training data for the LSTM model was that of 30-40 days, and ideally I would have liked to have far more training data as this would most likely have resulted in far better results. By the end of this time period I had successfully retrieved cryptocurrency related tweets and saved them to a CSV file.

Note: The code found in my 'SentimentAnalysis' script was written during this time



Acquire Twitter Data

Description

- Research the Twitter API (What are the ablilities & limitaions)
- Research Tweepy, which is a Python library for using the Twitter API
- Implement Tweepy calls to retrieve Twitter data
- Save retrieved data to CSV files.

Issues in this epic

100% Done

☑	~~FYP-20~~	Research the Twitter API (What are the ablilities & lim...	↑		DONE
☑	~~FYP-21~~	Research Tweepy, which is a Python library for using t...	↑		DONE
☑	~~FYP-22~~	Implement Tweepy calls to retrieve Twitter data	↑		DONE
☑	~~FYP-23~~	Save retrieved data to CSV files	↑		DONE

✓ Done

Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	3w logged
Priority	↑ Medium
Epic Name	Acquire Twitter Data

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

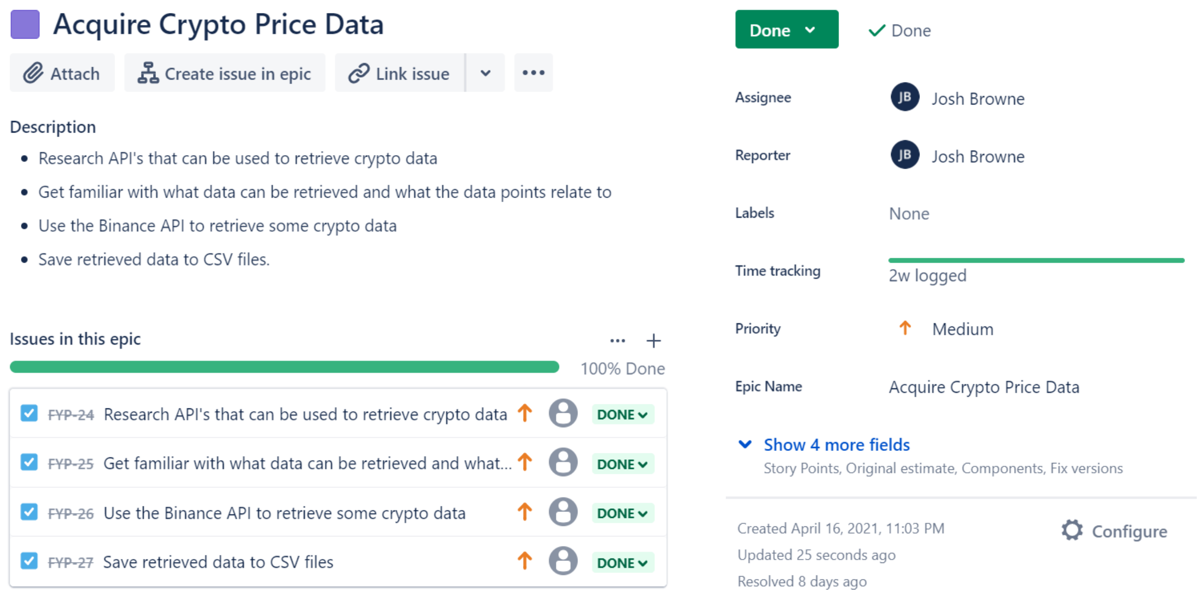
Created April 16, 2021, 11:03 PM
Updated 8 days ago
Resolved 8 days ago

February

The next part of my project to tackle was that of retrieving cryptocurrency price data. I investigated some resources in how this might be done. There are a wide selection of API's that can be utilised for this task, I chose to use the *Binance* API. To make a request to this API, you need to pass it the following parameters:

1. Symbol: The currency pairing we want data for.
2. Interval: The time step interval for each row of data.
3. Start time: The starting time we want the returned data to begin from.
4. End time: The time we want the returned data to finish on.

Once we pass these parameters correctly, we get a dataframe returned with data containing columns of datetime, open price, high price, low price, close price and volume. As I only needed the closing price data for each timestep, I saved the datetime and close price data to a CSV and discarded the other columns.



Acquire Crypto Price Data

Description

- Research API's that can be used to retrieve crypto data
- Get familiar with what data can be retrieved and what the data points relate to
- Use the Binance API to retrieve some crypto data
- Save retrieved data to CSV files.

Issues in this epic

100% Done

☑	~~FYP-24~~	Research API's that can be used to retrieve crypto data	DONE
☑	~~FYP-25~~	Get familiar with what data can be retrieved and what...	DONE
☑	~~FYP-26~~	Use the Binance API to retrieve some crypto data	DONE
☑	~~FYP-27~~	Save retrieved data to CSV files	DONE

✓ Done

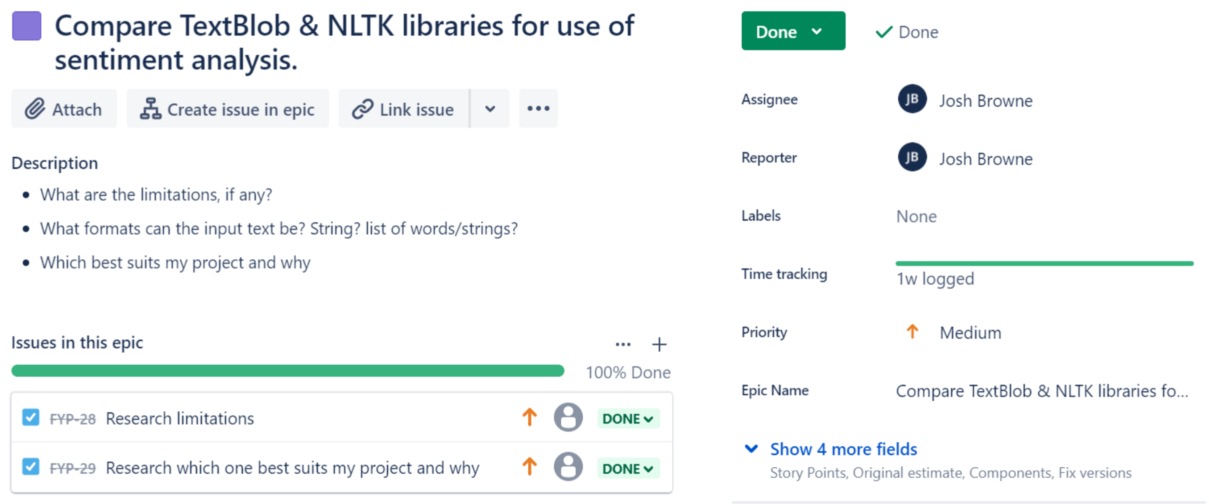
Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	2w logged
Priority	Medium
Epic Name	Acquire Crypto Price Data

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

Created April 16, 2021, 11:03 PM
Updated 25 seconds ago
Resolved 8 days ago

March

During this time I focused some time into comparing the TextBlob and Sentiment Analysis tools. I implemented both on junk data but found that since my twitter data will be cleaned and tokenized into a list of strings and that NLTK had the ability to perform sentiment analysis for this format, that NLTK was the preferable option. There were other reasons I found that indicated NLTK would better suit my use case, that being that the TextBlob lexicon was designed for review type texts and NLTK's was more general. I decided the more general approach would be safer as the data I would be using it on would be of a variety of text types and not strictly review based. Once the decision was made to use NLTK, I was able to write the scripts for retrieving the twitter data for each timestep in my cryptocurrency price database and perform sentiment analysis on the retrieved tweets to find the average sentiment and append this to my databases. The training data for the model was now ready for use.



Compare TextBlob & NLTK libraries for use of sentiment analysis.

Description

- What are the limitations, if any?
- What formats can the input text be? String? list of words/strings?
- Which best suits my project and why

Issues in this epic

100% Done

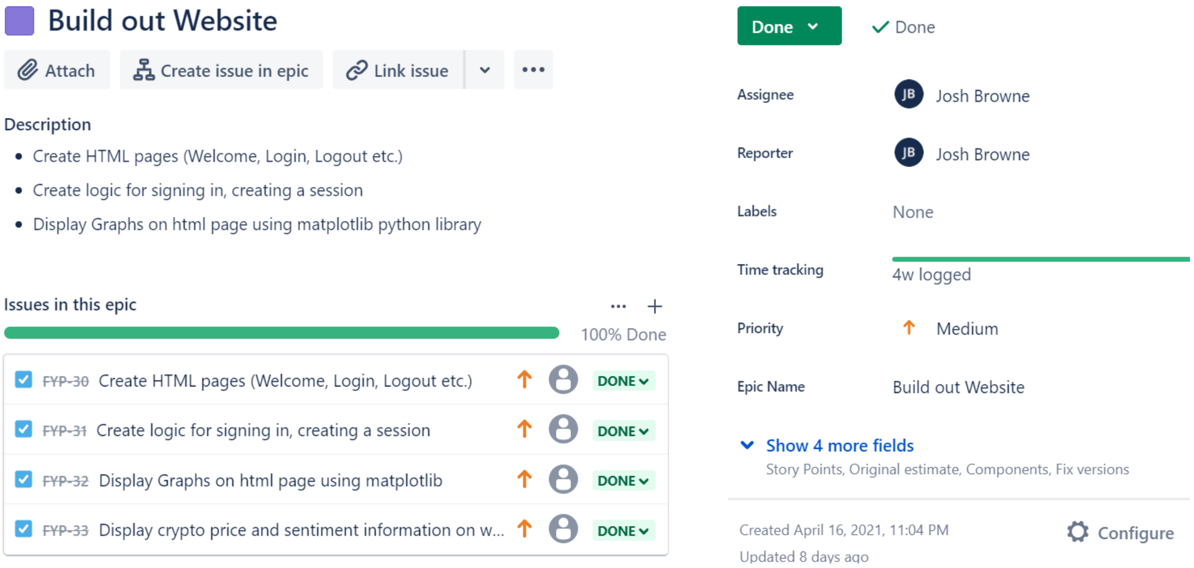
☑	~~FYP-28~~	Research limitations	DONE
☑	~~FYP-29~~	Research which one best suits my project and why	DONE

✓ Done

Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	1w logged
Priority	Medium
Epic Name	Compare TextBlob & NLTK libraries fo...

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

Other work carried out during this time was that of website development. I had the beginnings of my website HTML pages built out from work done over the christmas break. The work done at this time involved creating the Flask application and implementing the logic for signing up users and saving their details to a database for future logins. I also researched the MatPlotLib library and implemented the code needed for generating graphs as this would be how I will visualise my data.



Build out Website

Description

- Create HTML pages (Welcome, Login, Logout etc.)
- Create logic for signing in, creating a session
- Display Graphs on html page using matplotlib python library

Issues in this epic

100% Done

- ☑ ~~FYP-30~~ Create HTML pages (Welcome, Login, Logout etc.) DONE
- ☑ ~~FYP-31~~ Create logic for signing in, creating a session DONE
- ☑ ~~FYP-32~~ Display Graphs on html page using matplotlib DONE
- ☑ ~~FYP-33~~ Display crypto price and sentiment information on w... DONE

✓ Done

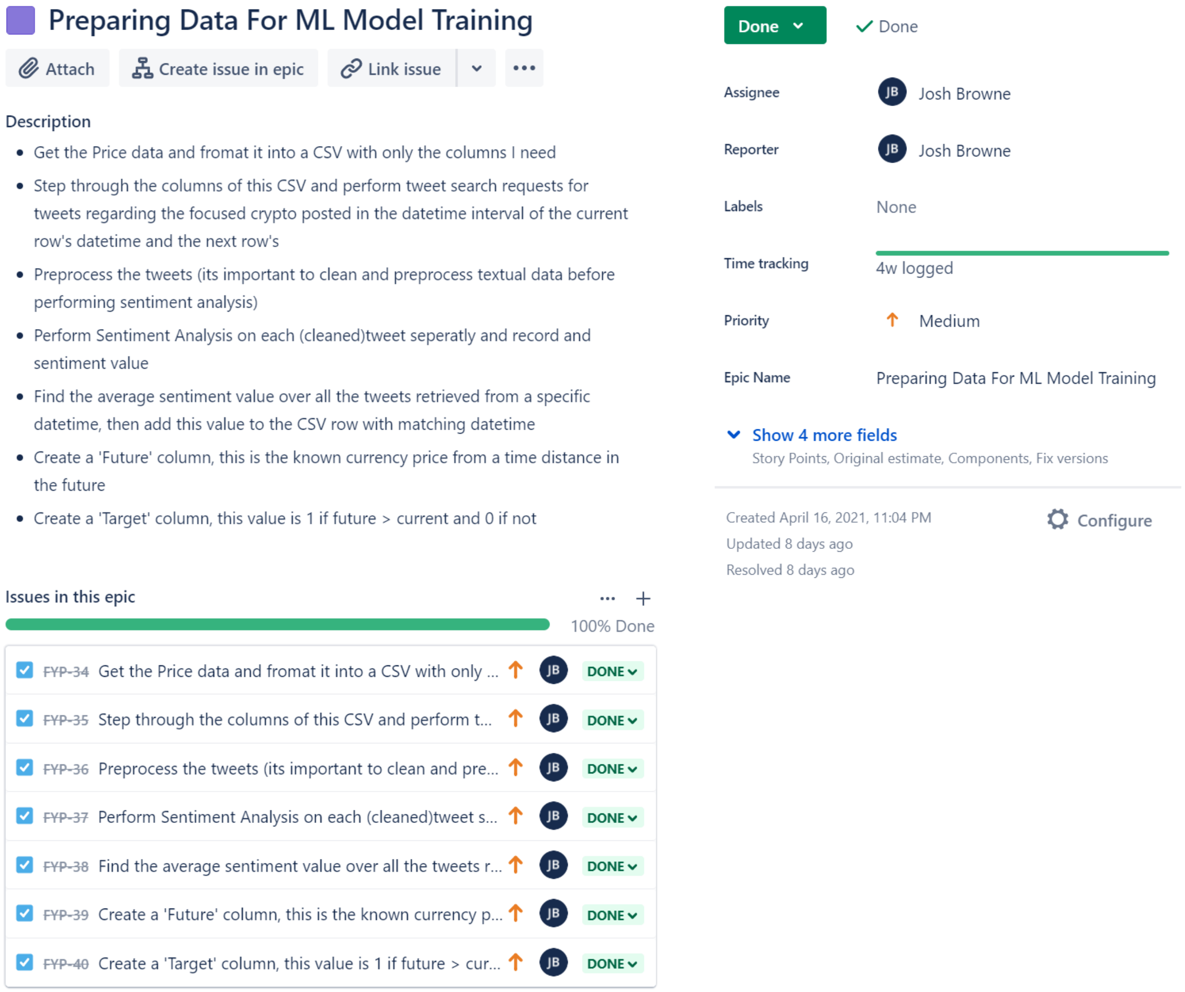
Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	4w logged
Priority	Medium
Epic Name	Build out Website

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

Created April 16, 2021, 11:04 PM
Updated 8 days ago

April

Now that I finally had acquired the data needed for training a model, it was time to develop the scripts to build my model. This involved sorting my data into batches and reshaping to fit the input parameters. Tensorflow models take in data scaled for 0-1 so I had to scale the price data down using MinMaxScalar. The work described here can be found in my “Model_Creation” script.



Preparing Data For ML Model Training

Description

- Get the Price data and fromat it into a CSV with only the columns I need
- Step through the columns of this CSV and perform tweet search requests for tweets regarding the focused crypto posted in the datetime interval of the current row's datetime and the next row's
- Preprocess the tweets (its important to clean and preprocess textual data before performing sentiment analysis)
- Perform Sentiment Analysis on each (cleaned)tweet seperatly and record and sentiment value
- Find the average sentiment value over all the tweets retrieved from a specific datetime, then add this value to the CSV row with matching datetime
- Create a 'Future' column, this is the known currency price from a time distance in the future
- Create a 'Target' column, this value is 1 if future > current and 0 if not

Issues in this epic

100% Done

☑	~~FYP-34~~	Get the Price data and fromat it into a CSV with only ...	↑	JB	DONE
☑	~~FYP-35~~	Step through the columns of this CSV and perform t...	↑	JB	DONE
☑	~~FYP-36~~	Preprocess the tweets (its important to clean and pre...	↑	JB	DONE
☑	~~FYP-37~~	Perform Sentiment Analysis on each (cleaned)tweet s...	↑	JB	DONE
☑	~~FYP-38~~	Find the average sentiment value over all the tweets r...	↑	JB	DONE
☑	~~FYP-39~~	Create a 'Future' column, this is the known currency p...	↑	JB	DONE
☑	~~FYP-40~~	Create a 'Target' column, this value is 1 if future > cur...	↑	JB	DONE

Done ✓ Done

Assignee	JB Josh Browne
Reporter	JB Josh Browne
Labels	None
Time tracking	4w logged
Priority	↑ Medium
Epic Name	Preparing Data For ML Model Training

Show 4 more fields
Story Points, Original estimate, Components, Fix versions

Created April 16, 2021, 11:04 PM
Updated 8 days ago
Resolved 8 days ago

I'd like to note that my first attempt at model training was to mark each row either a buy or sell time, and this was decided by if the future price (24 hours in the future) was greater than the current price it would be marked as buy, and sell if future price was below the current price. I made an attempt at training a model to give as output, a prediction on whether it thinks the future holds buy or sell opportunity. This effort fell short though as in the end it was not clear how to interpret the output data. Sadly, this first attempt at model creation was very time consuming. I eventually redesigned my model so that it would make a prediction on the actual future values instead.

Once I finally had my model working, I was able to feed it price and sentiment data pairings in the correct input format to generate predictions. I then plotted these predicted values on graphs against the real price data to show the accuracy of the models predictions. These graphs are displayed on my website.

Major Technical Achievements

I believe my major technical achievements in this project were that of successfully training an LSTM model for predicting future values based on time series price data paired with sentiment data. Although the resulting predictions from my models do not show high accuracy, I do believe that this may be down to some of the following factors:

1. My training data only spans over 30-40 days as the Twitter API only allows queries for tweets in the past 30 days with my Developer account level access. I believe that this is probably too short of a dataset and it probably means there is a high chance of bias, as 30 days is not enough to fully pick up on trends.
2. An important step would be to level out the rises and falls of price so that there are equal amounts of each, to eliminate certain bias in the training data. As my training data was not large enough, doing this resulted in a dataset far too short to train a model on.
3. The language people use on Twitter, especially when talking about cryptocurrency is quite unique and NLTK's lexicon does not account for the majority of these key phrases people use, this is partly why I had a big interest in using IBM's context specific sentiment analysis. In the end it was not feasible to create the dataset needed for this context specific sentiment analysis but I would think this method would yield far better results.

Although the results from my model may not assist a trader in making successful trades, the methods I have implemented are scalable and with some further work with the training data, could be used to generate better predictions.

I also think that there is some value to the sentiment data retrieved and that using graphs to plot the average hourly sentiment over a time and comparing that with the hourly price values over the same period of time, is interesting and possibly valuable information. By these methods, and further analysis of these graphs, similarities may be drawn from sentiment fluctuations which may often coincide with similar price fluctuations.

Project Review

What went well?

Overall I think the project went well, as I successfully finished what I originally set out to accomplish. That being the training of an LSTM model and using it to make predictions, the use of sentiment analysis on twitter data and the development of a web application to show all of my found data. As web development was not particularly a strength of mine, I was happy in the end with how it turned out.

What did not go well?

My first attempt at training a model did not go well because I tried to train the model to predict whether it was likely the future price would be above or below current price. This effort proved unusable as I was not sure how to interpret the output data. In short, I was expecting output values of either 1 or 0 to classify the price movement up or down. The results were giving values where I was not sure how to classify into these categories, so in the end I abandoned this method and went for my price prediction approach. I would say that the results from my research were not particularly groundbreaking in that the predictions from my model did not yield very high accuracy. I do think that this may be down to some of the reasons listed in the "Major Technical Achievement" section though, and that my methods are scalable and with more training data I think the predictions accuracy may likely increase.

If starting again, what changes would I make to my approach?

I would like to have begun collecting the sentiment data earlier so that by the end I would have acquired far more training data. With a larger dataset for model training, I would have been able to implement some slicing to the training data to lessen bias and possibly get better prediction results.

What advice would I give to someone beginning a similar project?

I would like to encourage another researcher to investigate IBM's context specific sentiment analysis tools. I would warn them that in order to do this, they would need a training set of minimum 500 data points of labeled example phrases in their chosen domain. I would like to see this investigated further as I believe current sentiment analysis tools struggle to give accurate scores of certain data types, particularly tweet data as language used on this platform is very unique and often contains sarcasm, which will probably not be picked up with other tools.

What did I think of my technology choices, retrospectively?

I think that my choices for technology tools in use for my project were in some cases the safer decision. I decided to use the more general lexicon library found in NLTK's VADER library. I would like to have attempted some of the other available tools for sentiment scoring as there may be other libraries which better account for Twitter slang terms.

Conclusions

In the end, my findings were that the NLTK library's sentiment analysis functionality may not yield very good results when applied to tweet data about cryptocurrencies. This is likely because the tweet format is very small, which generates a whole new dimension of problems like the use of slang, abbreviations, etc. This project reports on the exploration and preprocessing of data, transforming data into a proper input format and to classify the user's perspective via tweets into sentiment scores by building supervised learning models using Python and the NLTK library. In this project, I used mu;tiple training models to make predictions and tried the most appropriate model on my dataset. During the process, I learned the analyzing steps and using the tools in NLTK such as the stopwords corpus for removing from my data very useful. I wanted to see how well the given sentiments are distributed across the training dataset. To accomplish this task I wanted to see the common words in the dataset by plotting word clouds.

Overall I have developed a web app which displays sentiment and price data for various cryptocurrencies, and also displayed here are the results of my LSTM model for predicting price which is trained on this data.

Future Work

I expect there to be future research efforts in this field, where methods may involve the use of context specific sentiment analysis. Technologies such as IBM Watson already offer such capabilities and I would be interested to see if these implementations yield better results. From my research of previous efforts of similar goals to this project, I found that neural networks in most cases are the best method for training a computer to make price forecasts in this way. I would like to see if my methods would have better results if a far larger training dataset was used.

There are currently no up-to-date lexicon libraries which have sentiment values for common Twitter slang and more specifically cryptocurrency related slang terms. I wonder what results would be found if these slang terms were accurately accounted for.

References:

1. A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli, "Evolutionary dynamics of the cryptocurrency market," Royal Society Open Science, vol. 4, no. 11, November, 170623, 9 pages, 2017.
   View at: Publisher Site | Google Scholar | MathSciNet

2. G. Hileman and M. Rauchs, "Global Cryptocurrency Benchmarking Study," Cambridge Centre for Alternative Finance, 2017.
   View at: Google Scholar

3. Binance.com. (2017) Binance at: https://www.binance.com/.
4. Coinbase.com at: https://www.coinbase.com
5. Kraken.com at: https://www.kraken.com/

6. "2019 Crypto Hedge Fund Report" PwC & Elwood.
   View at: pwc-elwood-2019-annual-crypto-hedge-fund-report.pdf

7. S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, A peer-to-peer electronic cash system, Bitcoin, 2008.

8. Ethereum Foundation (Stiftung Ethereum). (2018) Ethereum. https://www.ethereum.org/.

9. Ripple. (2013) Ripple https://ripple.com/.
10. Iota (2018) at: https://iota.org/.

11. A. Baronchelli, "The emergence of consensus: a primer," Royal Society Open Science, vol. 5, no. 2, February, 172189, 13 pages, 2018.
    View at: Publisher Site | Google Scholar | MathSciNet

12. J. Barrdear and M. Kumhof, "The Macroeconomics of Central Bank Issued Digital Currencies," SSRN Electronic Journal.
    View at: Publisher Site | Google Scholar

13. P. Ciaian, M. Rajcaniova, and D. Kancs, "The economics of BitCoin price formation," Applied Economics, vol. 48, no. 19, pp. 1799–1815, 2016.
    View at: Publisher Site | Google Scholar

14. D. Enke and S. Thawornwong, "The use of data mining and neural networks for forecasting stock market returns," Expert Systems with Applications, vol. 29, no. 4, pp. 927–940, 2005.

View at: Publisher Site | Google Scholar

15. I. Madan, S. Saluja, and A. Zhao, Automated bitcoin trading via machine learning algorithms, 2015.

16. H. Jang and J. Lee, "An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information," IEEE Access, vol. 6, pp. 5427–5437, 2017.
View at: Publisher Site | Google Scholar

17. Nor Azizah Hitam, Amelia Ritahani Ismail, Faisal Saeed,
An Optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO) for Cryptocurrency Forecasting,
Procedia Computer Science,
Volume 163,
2019,
https://doi.org/10.1016/j.procs.2019.12.125.
(https://www.sciencedirect.com/science/article/pii/S1877050919321647)

18. Go, Alec, Lei Huang, and Richa Bhayani. "Twitter sentiment analysis." Entropy 17 (2009): 252.
View at: Google Scholar

19. Giachanou, Anastasia, and Fabio Crestani. "Like it or not: A survey of twitter sentiment analysis methods." ACM Computing Surveys (CSUR) 49.2 (2016): 1-41.
View at: Google Scholar

20. Liew, Jim, et al. "Cryptocurrency investing examined." The Journal of The British Blockchain Association (2019): 8720.
View at: Google Scholar

21. Wood, Gavin. "Ethereum: A secure decentralised generalised transaction ledger." Ethereum project yellow paper 151.2014 (2014): 1-32.
View at: Google Scholar

22. Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.
View at: Google Scholar

23. Ferro, Federico. "Implementing sentiment analysis to assess the perception of polarizing products: the Tesla Cybertruck case." (2020).
View at: Google Scholar

24. Alamoodi, Abdullah, et al. "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review." Expert systems with applications (2020): 114155.
View at: Google Scholar

25. Khare, Kaustubh, et al. "Short term stock price prediction using deep learning." 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT). IEEE, 2017.
    View at: Google Scholar

26. Torkil Aamodt, Predicting Stock Markets with Neural Networks-A comparative study.
    View at: Google Scholar

27. Hengjian Jia, Investigation Into The Effectiveness Of LongShort Term Memory Networks For Stock Price Prediction.
    View at: Google Scholar

28. Bollen, J.; Maoa, H.; Zeng, X. Twitter mood predicts the stock market. J. Comput. Sci. 2010
    View at: Google Scholar

29. Chen, C.-H.; Hafner, C.M. Sentiment-Induced Bubbles in the Cryptocurrency Market. 2019
    View at: Google Scholar

30. IBM Watson documentation. Found at:
    https://www.ibm.com/blogs/watson/2020/08/solving-common-challenges-in-sentiment-analysis-with-help-from-project-debater/