Twitter Sentiment Analysis on Cryptocurrencies

Jaskirat Singh, Omkar Pandit

Abstract

The two most appealing and flourishing technologies in the present world are machine learning and blockchain. In the past decade, research has grown exponentially in both these technologies and these are expected to bring in the next technological revolution. We aim to combine some of the use cases of both these technologies, to predict the topics that people talk about when they mention two most popular cryptocurrencies, i.e. Bitcoin and Ether. Moreover, to find the sentiment flow about these topics and cryptocurrencies among Twitter users. The sentiment can then be used to predict a rise or drop in the value of cryptocurrencies. A higher positive average sentiment value would probably imply a rise in the value as a people are talking good about them whereas, a higher negative average value probably imply that people are talking negatively about the cryptocurrencies over Twitter.

Introduction

Till now, there have been multiple machine learning research works on the blockchain technology since the advent of the blockchain. Both blockchain and deep learning started flourishing in the early years of present decade. Most of these research works have focused on the use of historical data to predict the prices of cryptocurrencies, such as the machine learning application in the stock market. Whereas, very little works have been done on finding out the relationship that might exist between the people who are trading cryptocurrencies and the topics that they relate to cryptocurrencies. Assuming a relationship between the two, we aim to build a model to predict these topics that Twitter users talk about for Bitcoin and Ethereum.

Our goal is to fetch tweets that have the words Bitcoin or Ethereum in it. This is because Bitcoin and Ethereum are still the most popular blockchains out there. The actual name of Ethereums cryptocurrency is Ether but it is not a popular name and people tend to use the name Ethereum to refer to the cryptocurrency as well. The fetching of tweets includes the text of the tweet, the location of the user (if available), the number of retweets and the location where the tweet was posted as well. The fetching will be done using Python BeautifulSoup package. As in every data mining project, our next step will be to do the data cleaning and refining. This includes cleaning of the text if there are any additional characters, changing the latitudinal-longitudinal addresses to country names and removing the duplicates (if any), which does not include removing the retweets. Then comes the main task to get the sentiment analysis for all the tweets by using Vader Sentiment.

Our goals is to implement multiple machine learning models before finding out the sentiment scores. This includes: Implementing a cluster analysis by implementing LDA to figure out the issues or attributes related to Bitcoin and Ethereum. Next task will be to find a relation between these issues/attributes to the overall sentiment score. Finding out the word frequency, by creating a TF-IDF dictionary using NLP, of all the words that appear in the tweets. Doing a lift analysis of some particular word pairs such as Bitcoin buy or Ethereum and sell etc. Future goals include classifying the user sentiment among USA and rest of the world (ROTW) and comparing it.

Related Works

Machine Learning for Predicting the Cryptocurrency Prices :

Predicting the price of Bitcoin using Machine Learning McNally, Roche, and Caton (2018) paper focuses on predicting Bitcoin prices in USD. Bitcoin Prices are sourced from the Bitcoin price index. The author built a prediction system by implementing Bayesian optimized recurrent neural networks (RNN) and a Long Short-Term Memory (LSTM) network. The paper Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin Greaves and Au (2015) analyzes the networks impact on overall Bitcoin price. So, the author investigated the way to predict Bitcoin prices based on Bitcoin network-based components. An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information Jang and Lee (2018) unfolds the effect of Bayesian neural networks (BNN) by analyzing time series data of the Bitcoin process. The authors compare the BNN model with other relevant linear and non-linear models before building the final prediction system for the Bitcoin process.

These projects were focused to predict the market prices of the cryptocurrencies based on the historical data. The project was aimed at building machine learning regression models that take in historical data as an input and predict the market rate for the test data as an output. Various machine learning algorithms such as Support Vector Machines, Decision Trees, Artificial Neural Networks etc. have been used in these projects. These are one of the first kind of projects done with the aim of combining blockchain and machine learning.

Automated Cryptocurrency Trading :

Using time-series and sentiment analysis to detect the determinants of bitcoin prices Georgoula et al. (2015) implemented time series analysis to understand the relationship between Bitcoin prices and economic dimensions, technological components, and overall opinion from tweets posted on Twitter. Sentiment analysis has been implemented using a machine learning algorithm called Support Vector Machines (SVM). Automated bitcoin trading via machine learning algorithms Madan, Saluja, and Zhao (2015) project is divided into two parts. In the first part, the authors used the data of 25 most relevant features related to Bitcoin price collected over the period of 5 years and build a prediction model. For the second part, they focused on only Bitcoin price data and examined the price changes after every 10 minutes to 10 seconds intervals. Volatility Analysis of Bitcoin Price Time Series Pichl and Kaizoji (2017) work focuses on the price of Bitcoin compare to standard currencies and their volatility over the period of 5 years. The author studied multiple currencies like USD, EUR, and CNY. They implemented feed forward neural network with 2 hidden layers using 10-day timestamp window.

These projects advanced upon the previous approach of predicting the cryptocurrency prices. The motivation was to build an online model to predict the prices of cryptocurrencies in real-time. Most of these projects have used deep learning algorithms such as ANN and RNN for the time series analysis and have done quite well as compared to the previous works. The applications have from this field have come out to be quite fantasizing and promising. These projects have done a tremendous job at combining the use of deep learning and blockchain.

Detecting/ Fighting Fraud in Cryptocurrency using Machine Learning :

To understand fraud in the Bitcoin network authors proposed a multifaceted approach and publish their work in a paper called A Multifaceted Approach to Bitcoin Fraud Detection: Global and Local Outliers Monamo, Marivate, and Twala (2016b). Machine learning algorithms like trimmed K-means and kd-trees have been used for fraud detection from both global as well as local level. In the end, a global outlier viewpoint outcasted other approaches. Anomaly detection in bitcoin network using unsupervised learning methods Pham and Lee (2016) focuses on anomaly detection because illegal works are often anomalous in financial networks. The goal of this approach is to find users and transactions that look most suspicious. So, for this authors implemented three unsupervised learning methods namely k-means clustering, Mahalanobis distance, and unsupervised Support Vector Machine (SVM) on a couple of graphs generated by the Bitcoin transaction network. Like multifaceted approach, Unsupervised learning for robust Bitcoin fraud detection Monamo, Marivate, and Twala (2016a) paper describes a way to implement anomaly detection in Bitcoin network using unsupervised machine learning algorithms. Authors implemented trimmed k-means as it is capable of simultaneous clustering of objects and fraud detection.

Another field under which a lot of research has been done is fraud/anomaly detection in cryptocurrency transactions using machine learning. Though it is very hard to alter a transaction once it is committed in the blockchain, the research has been focusing on ruling out the nodes that carry out fraudulent transactions. The motivation is to find anomalies in the transactions and then look out for fraud in the real-time data. This is done by providing the historical training data as input where the model (name some classification models) learns to classify legitimate or fraudulent transactions through labels, learns the parameters and values of feature variables for these set of transactions and try to classify the transactions in incoming test data as fraud or not. It is again a remarkable yet different approach at combining the use of blockchain and machine learning.

All the recent works have shown some great uses cases that come out because of the collaborative use of machine learning and blockchain, but our project is focused mainly on an idea that is not extensively researched upon by these works. We believe that getting to know of what topics do people relate with cryptocurrencies and with what sentiment is also important, and hence decide to carry out this project.

Methods

The project is divided into 2 main parts as follows, Understanding major topics/ areas/issues that, according to Twitter users, affect Bitcoing and Etheruem, Sentiment Analysis of Cryptocurrency related tweets. And this two parts are further divided into several steps.

Extract Data from Twitter :

Latest tweets about cryptocurrencies are extracted using the Twitter application account. Twitter API is used to download the tweets using personal API keys and secret tokens. Python packages like OAuth, urllib are used to set up a connection and access Twitter API. In the end, information about the Twitter user, location, and tweet text is extracted and stored in a CSV file. Around 10,000 tweets were extracted Joshi and Vidhya (2018).

Tweets Pre-processing and Cleaning :

As tweets are in a textual format, the pre-processing of the text data is a very important step to make given data ready for analysis. Tweets pre-processing, and cleaning is implemented to remove noisy and inconsistent data present inside them. Our goal is to clean the noisy data that is less relevant in obtaining sentiment of tweets such as special punctuation, numbers, characters Joshi and Vidhya (2018).

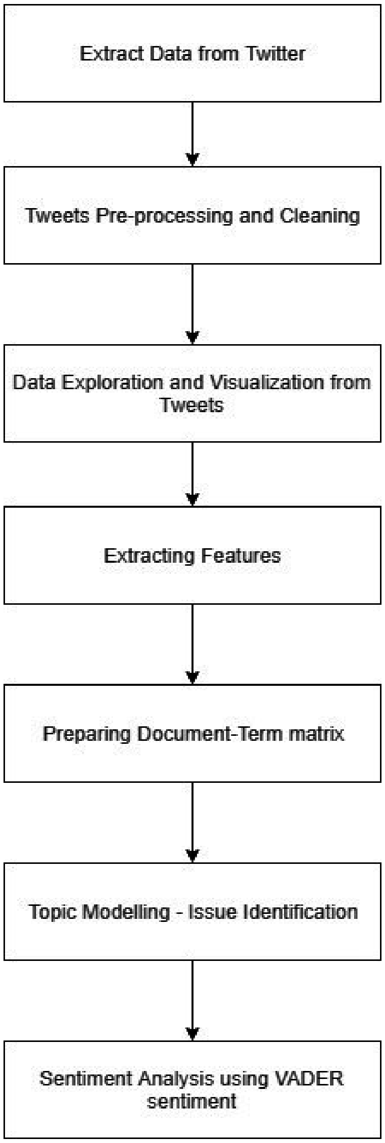
Extract Data from Twitter
Tweets Pre-processing and Cleaning
Data Exploration and Visualization from Tweets
Extracting Features
Preparing Document-Term matrix
Topic Modelling - Issue Identification
Sentiment Analysis using VADER sentiment

Figure 1: Flow Chart representing the entire process of the project

Data exploration and Visualization from Tweets :
Exploring and visualizing the data is important irrespective of input data type whether its in a textual or numeric format. This step really helped in gaining insights about the corpus/data. To understand that how the given sentiments are distributed across given set of tweets, we use Wordclouds. A wordcloud is one of the data visualization tools where the size of the appearance of the word depends upon its frequency of occurrence in a given dataset i.e. higher the frequency larger the size of an appearance of a word in a wordcloud Joshi and Vidhya (2018).

Extracting features from cleaned tweets using Data Modelling algorithms :
All machine learning techniques are designed to work with the numeric form of data. So, to analyze the pre-processed textual tweets data, it needs to be converted into a distinct numeric format. The data modelling techniques like Bag-of-Words, TF-IDF, and Word Embedding are used to do the conversion in order to obtain a highly accurate analysis of the sentiments Joshi and Vidhya (2018).

Preparing Document-Term matrix :
In text mining, each observation i.e. each tweet is considered as a document. A collection of all such documents together is called as a corpus. So, in order to run a mathematical model on a text corpus, the best way to do is to convert the text into a matrix representation. As Latent Dirichlet allocation (LDA) algorithm is implemented which focuses on repeating term patterns present in the entire Document-Term matrix, gensim library provided by corpora is used instead of traditional data modelling algorithms such as Bag-of-Words. Term dictionary is created for the whole corpus, where the unique term is assigned with an index. And at last this dictionary is used to convert a list of all documents present in corpus into Document Term matrix.

Running LDA model :
LDA model is trained on the previously created Document-Term matrix. Along with Document-Term matrix, LDA training also requires some additional parameters like Alpha and Beta hyperparameters, Number of Topics, Number of Topic Terms, and Number of Iterations/ Passes. Alpha and Beta are hyperparameters used to define document-topic density and topic-word density. Higher the value of alpha, the document can be composed of a greater number of topics and vice versa. And higher the value of Beta, the topic can be composed of a large number of words present in the corpus and vice versa. A number of topics parameter defines the exact number of topics to be extracted from the corpus. Different methods like Kullback Leibler Divergence Score are available to decide an optimal number of topics. A number of Topic Terms parameter defines how many numbers of terms can be composed in a single topic. The value of this parameter varies based on the given problem statement. Number of Passes defines a number of iterations allowed to LDA algorithm for clustering the given corpus into a predefined number of topics Li (2018).

Sentiment analysis using VADER sentiment :
VADER (Valance Aware Dictionary and sEntiment Reasoner) is a lexicon and a rule-based sentiment analysis tool used for this project, to understand the sentiment behind cryptocurrency related tweets. VADER is used because it is specially designed to measure the sentiments expressed in social media and Twitter is one of the majorly used social media platform to express individual opinions/views. The score is computed by adding the valance score defined by the lexicon for each word present in the given sentence/tweet. This score is normalized such that -1 defines extreme negative while +1 defines extreme positive. Generally, the output is categorized into three categories based on the value of a score. So, typical thresholds are, positive sentiment if the

score is greater than or equal to 0.5, neutral sentiment if the score is greater than -0.5 but less than +0.5, and negative sentiment if the score is less than -0.5 Calderon (2018).

Experimental Design

We aim to conduct multiple experiments to predict the sentiment analysis of Twitter users on two widely popular cryptocurrencies, Bitcoin and Ethereum. Not just that, we also aim at figuring out the reason behind the sentiment scores that we predict. We plan to achieve this objective by implementing topic modelling to figure out the issues that people are talking about over Twitter about Bitcoin and Ethereum.

Before carrying out different experiments, we will carry out data preprocessing tasks required to clean up the data. This includes data cleaning as well. The tweets scraped from Twitter contains some garbage values in itself such as the new line characters or the tab characters that we need to remove. Then we will be removing the blank values, if any in the dataset. Since we have unstructured data, we cannot replace the missing values with any substitute values such as mean or median. After removing the blank data points, we will remove the duplicate (if any) tweets as well.

The first task is to take the cleaned tweets as the input data and generate a lexicon using TF-IDF Ramos and others (2003). We do this with the aim of building a dictionary to carry out the natural language processing tasks on our data. Since the TF-IDF does not return a matrix, we also generated a Document Term Matrix (DTM). This DTM would be our input to the next experiment. Along with that, we are also calculating the word frequencies from our dataset, with the aim to match it to the results of our following experiments where we plan to predict the issues that the tweets depict are being related to Bitcoin and Ethereum.

Our experiment creates three Document Term Matrix, one for the entire corpus, second for the nouns in the tweets data and the last one for the for verbs. We made three LDA models each for nouns, verbs and the entire corpus with the aim of getting the keywords from the clusters made out of nouns and getting the actions out of clusters made out of verbs Wang et al. (2008).

Then we conducted topic modelling on the Document Term Matrix. We do this with the aim of creating different clusters of tweets from our entire dataset. To accomplish topic modelling, we have used Latent Dirichlet Allocation, which is a popular natural language processing algorithm Blei, Ng, and Jordan (2003). Finding the optical number of topics to make the clusters was a difficult task to achieve. Another task was to fetch a number of words that appeared more in the clusters, with the aim of depicting the issues out of it Kim and Shim (2014).

Our next experiment is to carry out sentiment analysis for our data. This is done with the aim of getting the sentiment values for each of the tweets in a cluster formed from the above experiment and get the average sentiment score of the cluster. We will take out the clean tweets from the clustered data and store them separately. We will pass these tweets as input to our python code that assigns the sentiment values to all the tweets using VADER Sentiment. The code returns four values against each document, a negative sentiment value, a neutral sentiment value, a positive sentiment value and a compound sentiment value Hutto and Gilbert (2014).

Our last task is to carry out lift analysis of the issues depicted by topic modelling with the keywords Bitcoin and Ethereum. If the lift analysis shows a strong connection between the issues and Bitcoin and Ethereum, it verifies the result given by topic modelling experiment.

Evaluation of topics clustered by LDA is very important as topic models cluster the given corpus of words into a number of topics in an unsupervised way. In general, it is hard to interpret the reasoning behind this clustering. There are multiple ways to judge the quality of topics. First, the easiest way is to use LDA visualization tool called pyLDAvis. We have used this tool and tried a various number of topics and just compared the results. The basic understanding is to avoid overlapping of word clusters as much as possible. Apart from visualization, we have also used coherence score measure to segregate between good versus bad topics and to estimate an optimal number of topics. There are a couple of methods like 'c_v' and 'UMass' are available to compute the coherence score. So, in this project, we have used 'c_v' method for calculating coherence score. To decide the optimal number of topics, we have calculated coherence score for a multiple number of topics and constructed a plot to display the relationship between the value of a coherence score and number of topics. Finally, based on human interpretability and a value of number of topics for which coherence score is maximum, considered as the optimal number of topics for considering issues related to Bitcoin and Ethereum cryptocurrencies Kumar (2018).

Experimental Results

The initial step returns a matrix of the number of words in the entire data set and their occurrences. The similar is done by word frequency analysis to verify the occurrence frequency of the dictionary created. Transforming the data into DTM is crucial to conduct topic modelling.

Since the tweets have a lot of short forms for longer words and hashtags, it becomes a bit difficult to get the accurate occurrence frequencies of the words which actually might be the same word.

For the experiment, we carried out the three LDA models for all the three input DTMs. From the LDA on the DTM made up of nouns, we found that the majority of the tweets were about Ethereum rather than Bitcoin. The second and third LDA models on nouns and verbs and overall

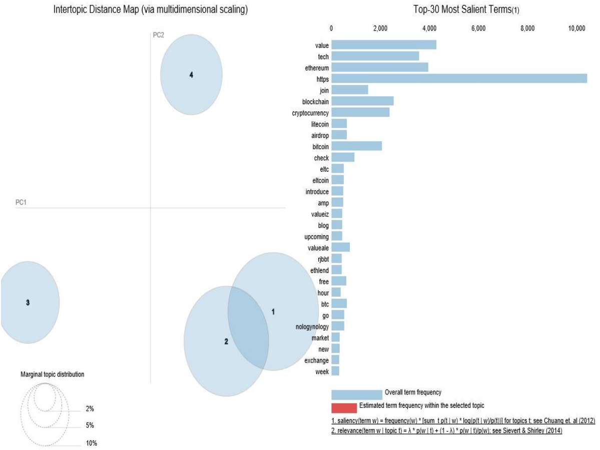

Figure 2: LDA Visualisation on whole corpus

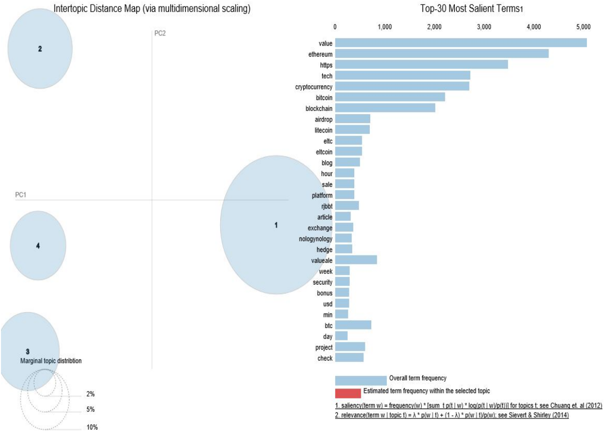

Figure 3: LDA Visualisation on Nouns

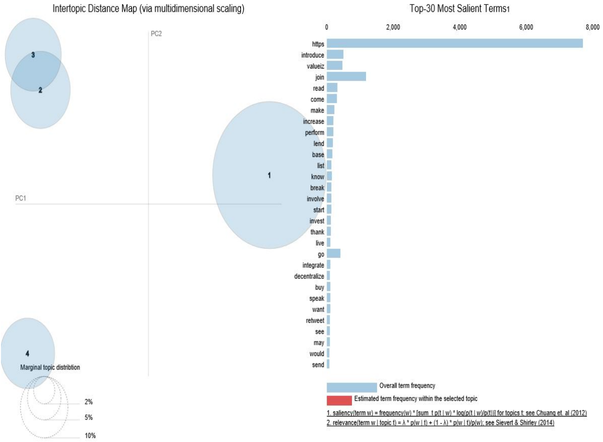

Figure 4: LDA Visualisation on Verbs

data helped in depicting the issues or attributes of the cryptocurrencies that the people are tweeting about. We tried generating multiple clusters and notice that they were overlapping each other. We choose to make 4 clusters because on making 4 clusters, we were able to get rid of any overlapping in the clustering as much as possible. LDA divides the whole corpus into 4 topics for further analysis. The clusters were understood by the high probability words that were allocated to the clusters. For each cluster, we decided to pick top 30 words, whose probability was higher of belonging to that cluster than other words. Upon getting these words and analyzing them we came up with some issues and attributes that people were talking about in relation to Bitcoin and Ethereum. These is included technology, value and security. Evaluation metrics like coherence scores and model perplexity. Both the metrics define how good the model is. Hence, lower the value of these measures better the quality of topics obtained.

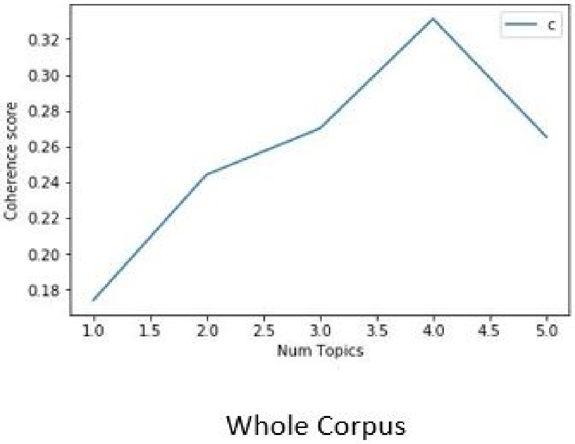

Figure 5: Coherence score vs Number of topics for LDA model

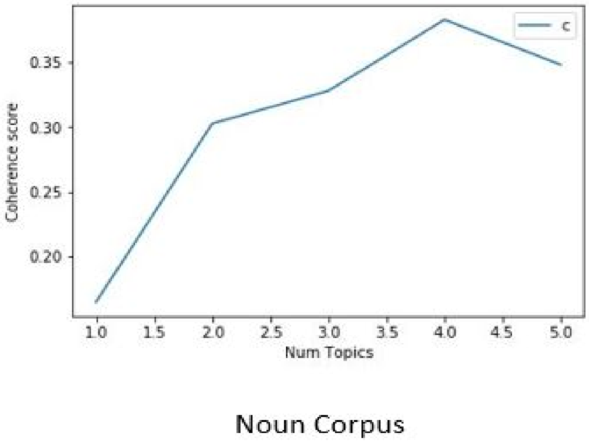

Figure 6: Coherence score vs Number of topics for LDA Noun model

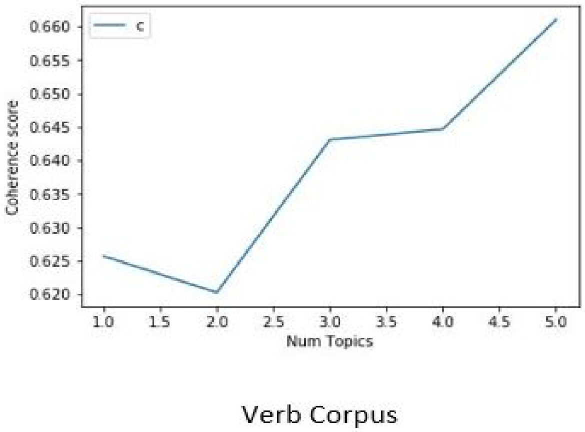

Figure 7: Coherence score vs Number of topics for LDA Verb model

The next experiment was to run a sentiment analysis of multiple divisions of data. We divided the data into multiple groups, where we had a significant number of tweets containing the keywords Bitcoin along with our issues and did the same for Ethereum. As we saw from our first experiment that the majority of tweets were about Ethereum and we had a significant number of tweets about Ethereum and all three issues whereas for Bitcoin, we had a significant number of tweets for only one of the issues. The results for the sentiment score of these groups of data are listed out in the table below and the graphs are shown as well.

Table 1: Final sentiments about Cryptocurrencies vs Issues

Cryptocurrency vs Issue	**Sentiment**
Bitcoin vs Security	Positive
Ethereum vs Security	Positive
Ethereum vs Technology	Positive
Ethereum vs Value	Strongly Positive

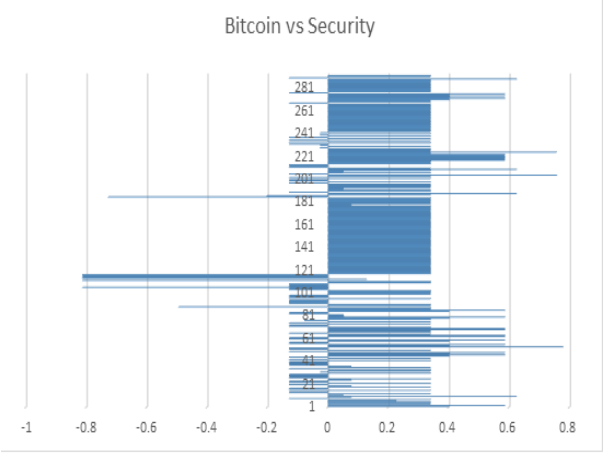

Figure 8: Sentiment Analysis of Bitcoin vs Security

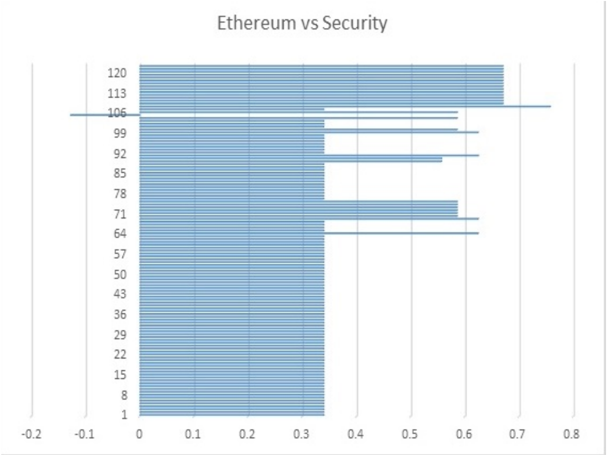

Figure 9: Sentiment Analysis of Ethereum vs Security

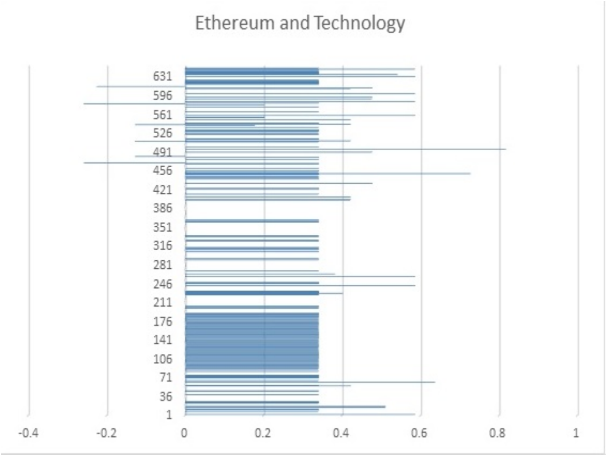

Figure 10: Sentiment Analysis of Ethereum vs Technology

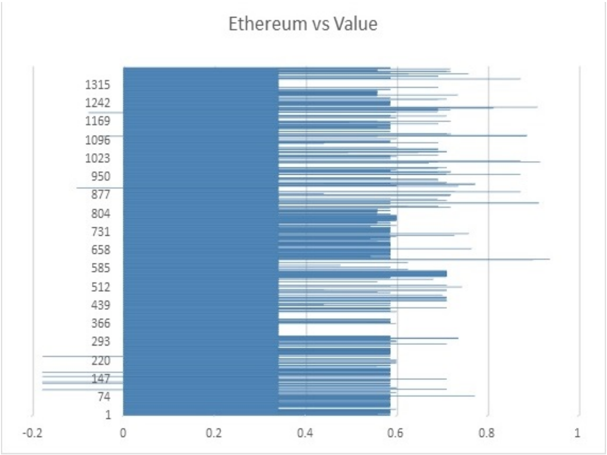

Figure 11: Sentiment Analysis of Ethereum vs Value

The last task was to carry out lift analysis for the word pairs Bitcoin and Security, Ethereum and Security, Ethereum and Value, Ethereum and Technology. The results for these pairs were making sense as they were justifying the clusters that we had generated in the above tasks. Our first and last task as such do not give out a presentable output but they were a crucial part in achieving our goal and also verifying our outcomes at multiple stages.

Conclusion

Our project focuses extensively on figuring out what people are talking about Bitcoin and Ethereum. Along with that, what are some of the issues that the Twitter users mention when they talk about these cryptocurrencies. As a future scopre, we assume that if overall, people are talking positively about them then it can probably be said that the value of these cryptocurrencies will rise and similarly if they have a negative sentiment towards it, the value can be expected to go down. Implementation of natural language processing lies at the heart of our project as we are dealing entirely with the textual data. Lastly, the main takeaway point from our project, is to use a combination of python packages such as BeautifulSoup and Sentiment analysis along with machine learning models to analyze user-generated unstructured data present over the social media to depict the public perspective about any topic.

Future Scope

We also plan to make country wise clusters and analyze the countrywide sentiment for both Bitcoin and Ethereum. We further aim at depicting the country wise issues or attributes that people tweet about Bitcoin and Ethereum. Also, if it could have been possible in the given time frame, we would have tried to build some additional models to improve upon our clustering algorithms and would want to get more accurate sentiment scores.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Calderon, P. 2018. Vader sentiment analysis explained.

Georgoula, I.; Pournarakis, D.; Bilanakos, C.; Sotiropoulos, D.; and Giaglis, G. M. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices.

Greaves, A., and Au, B. 2015. Using the bitcoin transaction graph to predict the price of bitcoin. *No Data*.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Jang, H., and Lee, J. 2018. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access* 6:5427–5437.

Joshi, P., and Vidhya, A. 2018. Comprehensive hands on guide to twitter sentiment analysis with dataset code.

Kim, Y., and Shim, K. 2014. Twilite: A recommendation system for twitter using a probabilistic model based on latent dirichlet allocation. *Information Systems* 42:59–77.

Kumar, K. 2018. Evaluation of topic modeling: Topic coherence.

Li, S. 2018. Topic modeling and latent dirichlet allocation (lda) in python.

Madan, I.; Saluja, S.; and Zhao, A. 2015. Automated bitcoin trading via machine learning algorithms. *URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan* 20.

McNally, S.; Roche, J.; and Caton, S. 2018. Predicting the price of bitcoin using machine learning. In *Parallel, Distributed and Network-based Processing (PDP), 2018 26th Euromicro International Conference on*, 339–343. IEEE.

Monamo, P.; Marivate, V.; and Twala, B. 2016a. Unsupervised learning for robust bitcoin fraud detection. In *Information Security for South Africa (ISSA), 2016*, 129–134. IEEE.

Monamo, P. M.; Marivate, V.; and Twala, B. 2016b. A multifaceted approach to bitcoin fraud detection: Global and local outliers. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, 188–194. IEEE.

Pham, T., and Lee, S. 2016. Anomaly detection in bitcoin network using unsupervised learning methods. *arXiv preprint arXiv:1611.03941*.

Pichl, L., and Kaizoji, T. 2017. Volatility analysis of bitcoin price time series.

Ramos, J., et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, 133–142.

Wang, D.; Li, T.; Zhu, S.; and Ding, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 307–314. ACM.