Sentiment Analysis on tweets for Bitcoin Price trend prediction

Chang Ti, Nan Jiang, Yuxin Mao, Tianhao Wu

Abstract:

1. Introduction

The development of cryptocurrency, digital coins that are not issued by any legal entity like fiat currencies (dollar or yen), (Abraham et al. 2018) in the past decade was rapid. In 2008 the first cryptocurrency was introduced to the world (Nakamoto 2008). Increasing in value gradually, the price of bitcoin skyrocketed to $290 in mid-2015 (Luther 2016) Eventually, The price of bitcoin climbed as high as $63,729.5, and it has fluctuated up and down unremittingly since then. However, bitcoin tumbled to $32,639, 7% of its value in one morning. It is not the only example, many other famous cryptocurrencies burgeoned followed by the growth of bitcoin. A massive number of them went down severely in a short period. Dogecoin was down 10% to $0.209537, and Ether was down 10.0% to $2,149.15. All happened in the same morning Bitcoin devalued, and thus all of the cryptocurrencies are considered very volatile.

The reason behind this abnormality and "chaos" is there exist some relationships between social media and the price of cryptocurrency, specifically bitcoin. As a result, similar to most commonly known currencies, cryptocurrencies are also affected by socially constructed opinions but on a more severe scale. (Georgoula et al. 2015) On January 29, 2021, Elon Musk changed the bio of his Twitter account to #bitcoin. This simple action has made the price of bitcoin rise by $6,000 (French 2021). Another example about how much Bitcoin/cryptocurrency is affected by social media is that on June 3, 2021, Musk simply tweeted a MEME (a humorous image, video, piece of text, etc., that is copied (often with slight variations) and spread rapidly by internet users.) about breaking up with bitcoin, and the price of bitcoin declined 5 percent as a consequence. Musk is only one example to show how the market of Bitcoin is deeply related to social media like Twitter.

Granted, grasping a precise understanding of cryptocurrency is not so easy(Abraham et al. 2018). It will still be helpful for prospective investors if there's a way to predict the trend of Bitcoin price by examining the sentiment of news and tweets. In this research, we make the following contribution: implementing machine learning models to gain a more considered understanding of the relationship between the sentiment of bitcoin-related tweets and bitcoin price.

2 Literature Review

Abraham et al. (2018) and Chen et al. (2020) both leveraged Google Trend Search Volume Index (SVI) and found a strong correlation between SVI and bitcoin price. The former research also included tweets volume and found that sentiment of tweets was not a determining indicator when bitcoin price was falling. However, Stenqvist and Lonno (2017) have analyzed sentiment fluctuation of over 2 million bitcoin-related tweets and gained a model predicting bitcoin price with 79% accuracy. Raju and Tarif (2019) have also shown that bitcoin price is predictable using sentiment analysis of bitcoin-related tweets. They implemented machine learning methods like Recurrent Neural Networks (RNNs) with Long Short Term Memory LSTM) and standard method ARIMA, and both models achieved high accuracy. This further reinforces that sentiment analysis can be an indicator of bitcoin price.

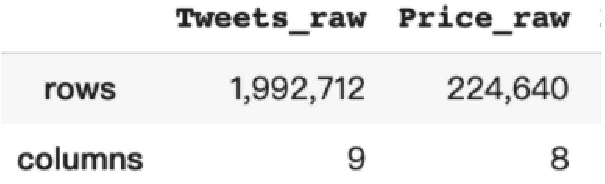
Ante (2021) has revealed that Elon Musk, who has 44.7 million followers on Twitter, has an extremely powerful influence on the cryptocurrency market and has caused a few abnormal trading volumes. They have also discovered abnormal returns of up to 18.99% for Bitcoin and 17.31% for Dogecoin after Musk's tweets. This research shows that magnates like Musk exert more impacts on the market. However, Mai et al.(2018) implemented a VECM (Vector Error Correction Model) to study the relationship between social media and the monetary value of bitcoin and revealed that the silent majority is the group who exerts significant effects, that is to say, users who are less active are the ones who dominate the bitcoin price.

3 Data

Our research leverages both Natural Language Processing (NLP) and Machine Learning (ML) to study the correlation between the bitcoin price trends and social media sentiment. We gathered two datasets one of which is the bitcoin price collected for every minute from 2016 to 2020.

The other dataset is bitcoin-related tweets from 2016 to 2019 from Kaggle (https://www.kaggle.com/). The tweets dataset contained 1992712 columns and 9 rows with feature number of users like the tweets etc. The price contain 224,640 rows and 8 columns with features like opening price or closing price.

Fig 1. Statistics about the datasets

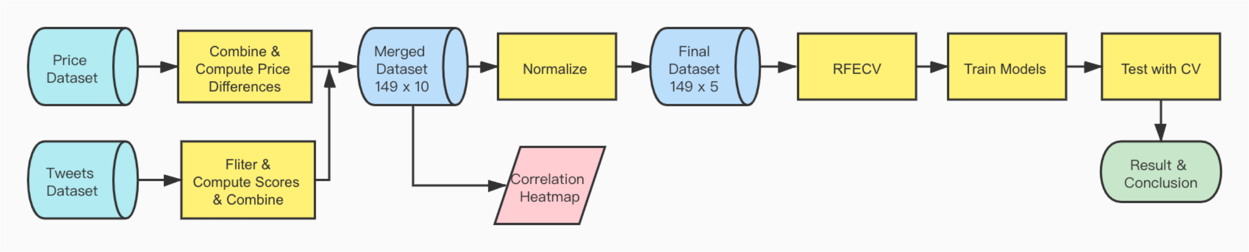


	Tweets_raw	Price_raw
rows	1,992,712	224,640
columns	9	8

4 Methodology

The chapter is organized according to the chronology of the experiments: data preprocessing and description for prediction model and its evaluation. We employed two approaches to measure sentiment of tweets. We computed the polarity and subjectivity using TextBlob (https://textblob.readthedocs.io/en/dev/) as the first approach, and computed sentiment scores using the vaderSentiment API (https://github.com/cjhutto/vaderSentiment) as the second approach. Our goal is to find a better way of analyzing tweets for bitcoin price prediction.

Flow Chart

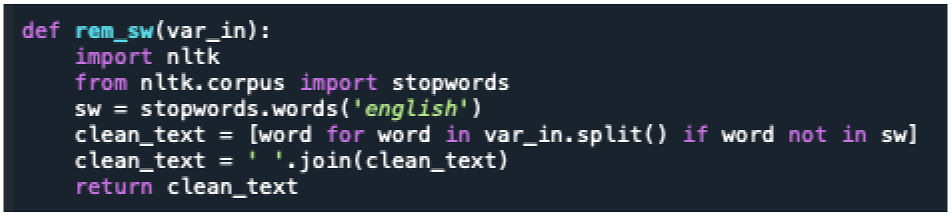


4.1 Data cleaning

In the price dataset specifically, we selected the open price of the day and the price difference between day t and day t-1 to day t-6. Then we added a column called label that generated by minus the current day price from the next day's, and transferred the exact number to -1(going down), 0(stay the same), and 1(going up). This way, we have achieved the goal of predicting the future price trend.

For the tweets dataset, we filtered meaningless text (e.g. hashtags, URLs, and certain special characters) and discarded irrelevant columns (e.g. user name and ID number), also columns with missing values. The raw dataset contained more than 1.3 million tweets and after cleaning, 1221912 tweets posted in the first 149 days of 2019 are retained. We cleaned the text and kept only English characters, and removed stop words which are words containing no specific meaning. Finally, we applied a dictionary to check on the text and manually remove words like "bt" or "co", characters only serving as noise for further analysis.

Example function that cleaning the stop words



```
def rem_sw(var_in):
    import nltk
    from nltk.corpus import stopwords
    sw = stopwords.words('english')
    clean_text = [word for word in var_in.split() if word not in sw]
    clean_text = ' '.join(clean_text)
    return clean_text
```

4.2 Data Preprocessing Vader

Valence Aware Dictionary and sEntiment Reasoner (VADER) is developed by Hutto and Gilbert(Hutto, 2014). It is commonly used for calculating the sentiment intensity of the text in the social media domain. Vader calculated a compound sentiment score ranging from -1.0 (negative) and 1.0 (positive) according to the lexicons.

4.2.1 Sentiment score and Aggregating

Every processed tweet is passed in and calculated individually by Vader(See Section 2.2.1), and each tweet row in the data set is tagged with an additional individual sentiment score. Average scores were calculated when combining tweets from the same day, and we also track down the number of tweets per day as “Volume” along the way. Afterward, we merged the processed price and tweets datasets according to date.

4.2.2 Additional Cleaning

To further normalize the data in order to get a better result, we first applied the simple return method for the t-1 to t-6 and observed that the numbers become too minimal.

Simple Return Formula

$$Simple\ Return = \left\{\frac{Present\ NAV - Starting\ NAV}{Starting\ NAV}\right)$$

Note: Nav stands for price t in this case

As a consequence, we decide to apply a z-score for every column in the merged table except the label to finally normalize the data.

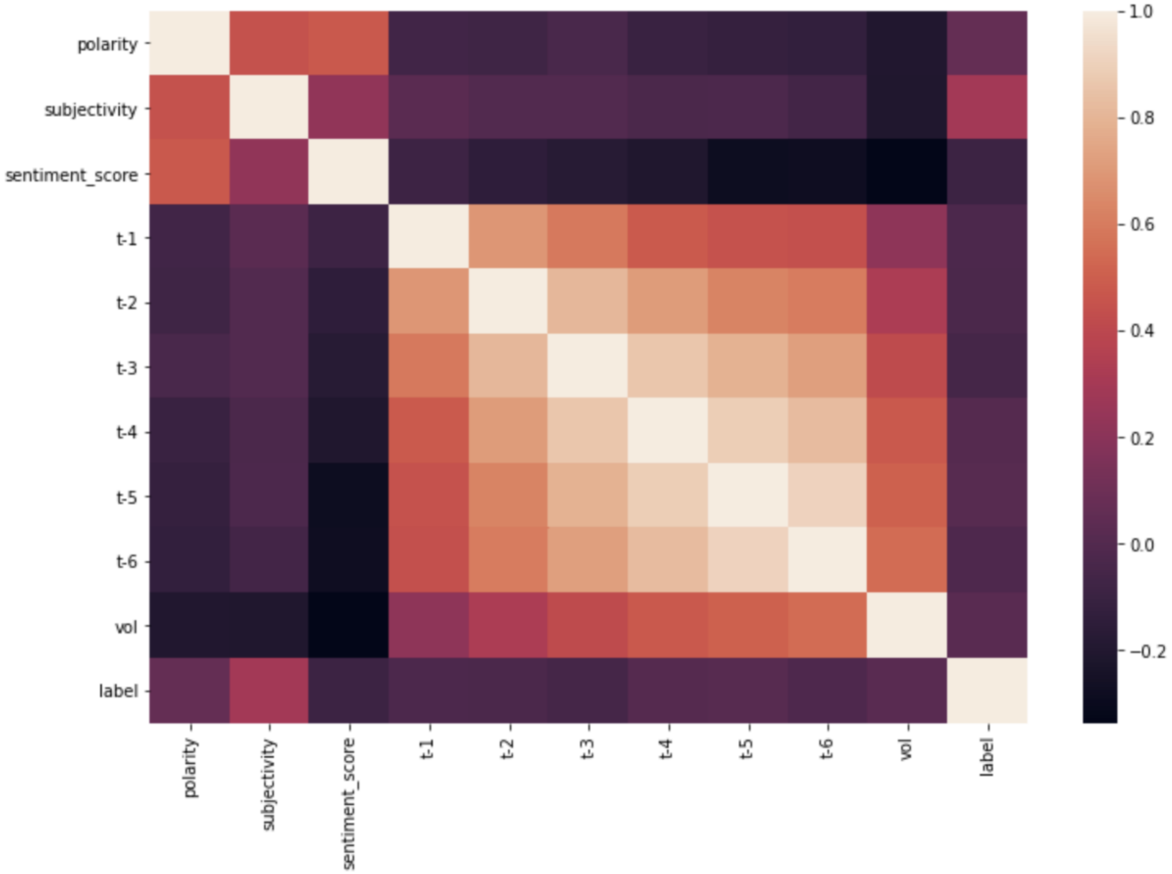
Z-Score Formula

$$Z = \frac{x - \mu}{\sigma}$$

4.3 Data Preprocessing TextBlob

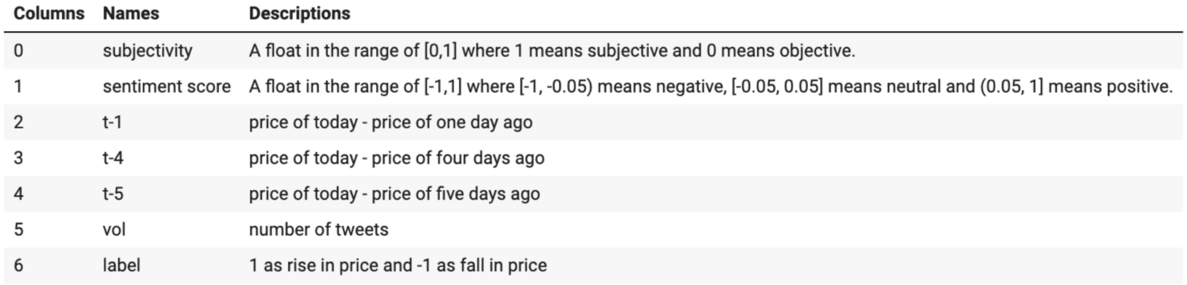
We plotted out a correlation heatmap to observe the correlations among the 11 features. It does not show a strong correlation between labels and other features, but it is observed that vol is correlated with price difference where the correlation gets higher from t-1 to t-6.

Figure 1. Correlation heatmap of all features.



In order to achieve an efficient modeling process, we employed Recursive Feature Elimination with Cross Validation (RFECV) to eliminate features that have less influence on bitcoin price. This resulted in maintaining 6 features after this feature engineering step..

Figure 2. Description of the selected features and label

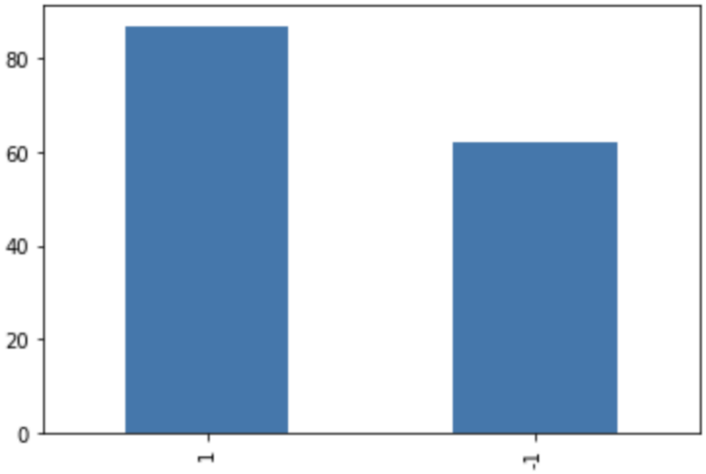


Columns	Names	Descriptions
0	subjectivity	A float in the range of [0,1] where 1 means subjective and 0 means objective.
1	sentiment score	A float in the range of [-1,1] where [-1, -0.05) means negative, [-0.05, 0.05] means neutral and (0.05, 1] means positive.
2	t-1	price of today - price of one day ago
3	t-4	price of today - price of four days ago
4	t-5	price of today - price of five days ago
5	vol	number of tweets
6	label	1 as rise in price and -1 as fall in price

Note: This table represents the features used in the modeling process after feature engineering

Since the variables have significantly different ranges and most of them do not follow normal distribution in a strict way, we normalized them by using the MinMaxScaler function in Scikit-learn. The features range from 0 to1 after normalization.

Figure 3. Distribution of label values.



4.3 Model Training

We trained eight models including Dummy Classifier, Logistic Regression, SVC, Decision Tree, Random Forest, K Nearest Neighbors, Gaussian Naive Bayes and Voting Classifier with two different groups of features respectively. Two groups were trained in the same way.

Since Dummy Classifier totally relies on the majority of classes in the training set, which is not a smart algorithm for machine learning scientists but a common method used in real life, we will take it as the baseline for our research. Logistic Regression runs fast but is usually not the algorithm performing well. However, we took this algorithm into account as we do not have many features and we predict that the class distribution is close to linear, thus, Logistic Regression should work well on this problem. Also, as what we have noticed in correlation heatmap, the features do not show much correlation with each other. Vol and price differences in previous days are correlated but feature vol was eliminated in the feature selection process. Naive Bayes performs well when features are independent, so we believe NB should be a good choice here.

In order to make the research thorough, we trained seven models and the Voting Classifier for the majority based on these seven models.

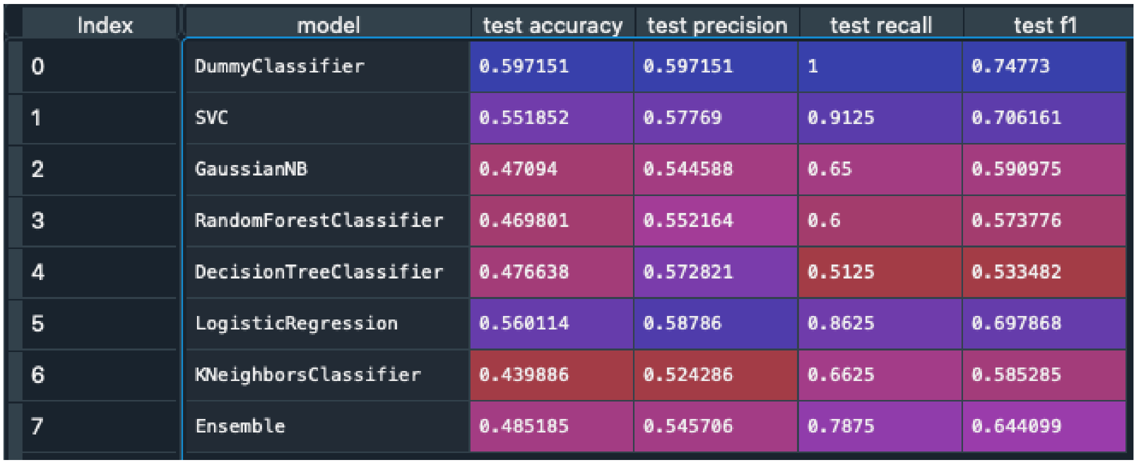
5 Result

Two approaches came to similar results where SVC, Logistic Regression and Naive Bayes are the three models with highest F1 score.

5.1 Vader Sentiment Analysis

Data sets for both price and tweets started on Jan 16 to May 29, 2019. The final table combined 134 days of Bitcoin related tweets and price. Table 1 shows the prediction performance of all seven models, and ensemble score of them(See section 3.1).

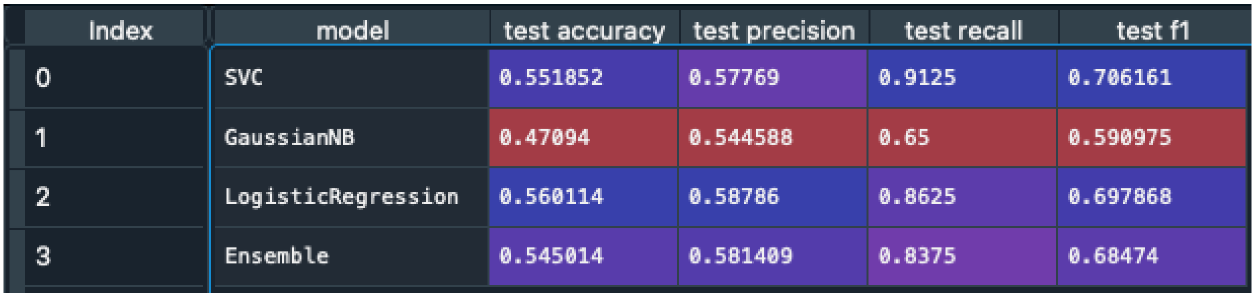
Table 1



Index	model	test accuracy	test precision	test recall	test f1
0	DummyClassifier	0.597151	0.597151	1	0.74773
1	SVC	0.551852	0.57769	0.9125	0.706161
2	GaussianNB	0.47094	0.544588	0.65	0.590975
3	RandomForestClassifier	0.469801	0.552164	0.6	0.573776
4	DecisionTreeClassifier	0.476638	0.572821	0.5125	0.533482
5	LogisticRegression	0.560114	0.58786	0.8625	0.697868
6	KNeighborsClassifier	0.439886	0.524286	0.6625	0.585285
7	Ensemble	0.485185	0.545706	0.7875	0.644099

We selected three models with the best performance in f1 score to proceed the experiment. However, Dummy Classifier is excluded cause the test recall is 1 and we found it suspicious. Result shows in Table 2.

Table 2



Index	model	test accuracy	test precision	test recall	test f1
0	SVC	0.551852	0.57769	0.9125	0.706161
1	GaussianNB	0.47094	0.544588	0.65	0.590975
2	LogisticRegression	0.560114	0.58786	0.8625	0.697868
3	Ensemble	0.545014	0.581409	0.8375	0.68474

Logistic regression performs the best in terms of test accuracy for 56.01% and test precision for 58.79%. Support Vector Machine(SVC) garnered the highest score in test recall and F-Score for 91.25% and 70.06%(See section 3.2). Except Gaussian Naive Bayes, all other three perform similarly with range smaller than 2% for test accuracy test precision and F1-score, 5% for test recall. We use F1 score as the indicator, so the Support Vector machine is the best performing model among all seven.

5.2 TextBlob

All of these three models reach an F1 score of around 0.73 while F1 score of Dummy is 0.595. Logistic Regression and SVC reach 0.965 and 1.000 for recall in respect, and 0.591 and 0.583 for precision respectively. High recall rate indicates that these two models can capture almost all days with price rise, but relatively low precision reduces the value of capturing all price rises. We have 149 instances in total and around 58.4% (which is even higher than precision of SVC) of them are price rise, thus, if we predict all instances in the test set as positive (price rise), we will obtain recall of 1.00 and precision of 0.584, which is better than SVC. 0.73 is a high score for F1 of a price prediction case, but considering the low precision, it is not convincing to use these two models in real-life investment. Naive Bayes, however, shows higher potential to be used in investment. It has the highest scores for accuracy, precision and recall among all models. Precision is 0.642 and recall is 0.864, which indicates that the NB model captures most price-rise-days and 64.2% of days that are predicted as rise shows a real rise. Therefore, the NB model is the most recommended model in this case.

6 Discussion and Future Work

Due to time limitation, we didn't collect our database but used the one available on Kaggle. In addition, the data we used was in 2019. Besides, Bitcoin is a heated topic all over the world, so the data set contain many tweets in other languages like Japanese or Spanish. We didn't get a chance to translate those tweets into English. We weren't able to process the whole dataset as well cause of hardware limitations, and we were only able to extract 149 consistent date of data out of the original one. Due to the size of the datasets, we ran into a problem of data splitting. We noticed that when random state of train_test_split (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) by scikit-learn was 42, all performance scores are extremely high. This is not frequent but could happen especially when the dataset is not large enough, thus, we employed 5-fold cross validation to avoid this kind of bias in the later experiment.

For future improvement, a more recent database would be favorable. If we are able to get access to a more powerful computer, it’s likely that we will increase the size of our dataset, which means to take more tweets into account. Also, applying various weight to tweets according to people’s wealth or social status, and their influence on the crypto currency world or the society would be an interesting idea to pursue forward. Another possible future direction based on what we have done is to study the correlation of the number of tweets and price differences. It is noticeable that there is a correlation, but the reason behind it and the relation between bitcoin price and investors’ other behaviors on social media is unknown and worthy of further research.

Reference:

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1.

Ante, L. (2021). How Elon Musk's Twitter Activity Moves Cryptocurrency Markets. *Available at SSRN 3778844*.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. Foundations and TrendsR in Information Retrieval, 2(1–2):1–135, 2008.

Bouoiyour, J., Selmi, R., Tiwari, A. K., & Olayeni, O. R. (2016). What drives Bitcoin price. Economics Bulletin, 36(2), 843-850.

Burnie, A., & Yilmaz, E. (2019). Social media and bitcoin metrics: which words matter. *Royal Society open science*, *6*(10), 191068.

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Colianni, S., Rosales, S., & Signorotti, M. (2015). Algorithmic trading of cryptocurrency based on Twitter sentiment analysis. *CS229 Project*, 1-5.

French, J. J. (2021). # Bitcoin,# COVID-19: Twitter-Based Uncertainty and Bitcoin Before and during the Pandemic. *International Journal of Financial Studies*, *9*(2), 28.

Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D., & Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *Available at SSRN 2607167*.

Guégan, D., & Renault, T. (2021). Does investor sentiment on social media provide robust information for Bitcoin returns predictability?. *Finance Research Letters*, *38*, 101494.

Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.

Jang, H., & Lee, J. (2017). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *Ieee Access*, *6*, 5427-5437.

Luther, W. J. (2016). Bitcoin and the future of digital payments. *The Independent Review*, *20*(3), 397-404.

Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. (2018). How does social media impact Bitcoin value? A test of the silent majority hypothesis. *Journal of management information systems*, *35*(1), 19-52

Matta, M., Lunesu, I., & Marchesi, M. (2015, June). Bitcoin Spread Prediction Using Social and Web Search Media. In *UMAP workshops* (pp. 1-10).

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, 21260.

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018, October). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)* (pp. 128-132). IEEE.

Pano, T., & Kashef, R. (2020). A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing*, *4*(4), 33.

Raju, S. M., & Tarif, A. M. (2020). Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis. *arXiv preprint arXiv:2006.14473*.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, *21*(6), 589.

Yermack, D. (2015). Is Bitcoin a real currency? An economic appraisal. In *Handbook of digital currency* (pp. 31-43). Academic Press.