USING SENTIMENT ANALYSIS ON TWITTER TO PREDICT THE PRICE FLUCTUATIONS OF SOLANA

SAMUEL VAN WIJHE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

269339

COMMITTEE

dr. Mirjam de Haas
dr. Eriko Fukoda

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 14, 2022

ACKNOWLEDGMENTS

I first and foremost want to thank the persons providing me with the used libraries for my research. The developers of the VADER algorithm, Twint library as well as the developers of the time series models I utilized. On top of that, I want to thank Nachiketa Hebbar a Youtuber who helped me get a better understanding of the inner workings of time series models. I would like to extend my sincere thanks to my supervisor Mirjam de Haas, who despite having different expertise than my research topic, provided me with excellent guidance and helped me better understand the structure and tendencies of the report itself. I am also grateful to Jishnu and Lindsey for working alongside me to keep me motivated. Without them, I would have probably been under a severe time pressure. Lastly, I want to thank Saifedean Ammous, writer of the book, The Bitcoin Standard. This book made me interested in cryptocurrency in the first place. The book gives a clear understanding of the rise of decentralized, apolitical, free market alternative to central banking systems.

USING SENTIMENT ANALYSIS ON TWITTER TO PREDICT THE PRICE FLUCTUATIONS OF SOLANA

SAMUEL VAN WIJHE

Abstract

Cryptocurrency and blockchain are some of the most debated new technologies in the past decade. They have changed the functioning of currencies in a significant way. One of the reasons for this is the enormous returns in a short period of time. Twitter and other social media platforms provide a great way to express opinions and share thoughts with others which could drive people to certain actions. This paper tests whether sentiment analysis is an useful tool for predicting price fluctuations in an alternative cryptocurrency called Solana by using the sentiment expressed on Twitter. By using Twitter data to derive sentiment and historical price data to forecast future values. The data set comprises of 31,060 tweets gathered by using the TWINT library consisting of tweets containing the dollar sign "SOL". The collected tweets about Solana are converted to polarity scores that represent the positivity/negativity of the tweet. This is done by a lexicon and rule-based sentiment analysis tool called VADER (Valence Aware Dictionary and sEntiment Reasoner). After computing the polarity scores are added up to represent a collective sentiment per hour. Functioning to be the predictor variable for the cryptocurrencies' price.

An univariate time series model's prediction is then compared with a multivariate time series (one which includes the sentiment) to test whether performance increases by making use of sentiment scores obtained from the collected tweets. By evaluating both models, the findings of this research conclude that sentiment alone with historical price data does not enhance the model's performance. Thus, it can be clearly stated that other features are required to develop a good forecasting model for Solana.

1 INTRODUCTION

Cryptocurrencies are digital representations of value that exist purely in electronic form based on blockchain technology. It is a revolutionary technology based on decentralized and cryptographic principles that function as an intermediary in digital transactions. It makes the roles of third parties obsolete. The blockchain technology that underpins it, takes advantage of peer-to-peer networks to create a shared and trusted ledger of transactions. Blockchain, as a distributed ledger, records transactions as an irrevocable timestamped digital block. Using both senders and receivers. The network does not make use of a centralized authority. Only members can validate transactions among themselves. This novel method of distributed data storage and management keeps track of all activity in real-time.

Since the beginning of Bitcoin, social media has been the platform for cryptocurrency information. Researchers have used postings from forums like Reddit and Coindesk to apply sentiment analysis and estimate price fluctuations in Bitcoin. Wołk (2020) investigated how social media affects cryptocurrency pricing. On this subject, Google trends and Twitter were used to predict short-term pricing of cryptocurrencies. The findings indicate that price swings of cryptocurrencies are substantially influenced by social media sentiment. In a study conducted by Mai, Shan, Bai, Wang, and Chiang (2018) findings indicate that "social media sentiment is an important leading indicator of future Bitcoin price swings."(Mai et al., 2018, p. 42). Most experts emphasize the importance of social media and news sentiment in predicting Bitcoin prices of trading over short periods, hourly to daily basis (Bollen, Mao, & Zeng, 2011; Mai et al., 2018; Wołk, 2020).

Whilst news has a significant impact on the crypto market, public opinion or attitude may be just as important. It is known from psychological research that emotions play an important role in human decision-making (Kemdal & Montgomery, 2002). Behavioural finance has revealed more evidence that sentiment and mood play a key role in financial decision making (Nofsinger, 2005). As a result, there is a reasonable assumption to believe that public sentiment and attitude can influence stock market prices just as much as the news.

A study conducted by Pagolu, Reddy, Panda, and Majhi (2016) used micro blogging to predict stock prices, as it validly reflects people's opinions and feelings about current events. They applied sentiment analysis with the use of supervised machine learning principles to study the relationship between tweet sentiments and the movements of Microsoft's stock price. According to the researchers, the study shows that there is a definite link between the fluctuations of stock values and the public opinion on Twitter. Illustrating that there is "a strong correlation exists between the rise and

falls in stock prices with public sentiments in tweets" (Pagolu et al., 2016, p. 1349). The authors used a classification model that was able to predict whether the price of a stock would increase or decrease. While this seems promising for stock price prediction, in the light of cryptocurrencies this is not always the case. In cryptocurrency trading commission fees are something that should be accounted for. While the classification model can predict if the prices go up, it does not state by how much. Therefore, the classification model could predict an increase, but if the increase is less than the transaction fees it is not a proper tool for investment advice.

This research focuses on using sentiment analysis to increase a models predictive performance on an alternative cryptocurrency called "Solana" by making use of sentiment on Twitter data. The platform is useful to generate a vast amount of sentiment upon data analysis. By making use of the VADER (Valance Aware Dictionary for Sentiment Reasoning) model the opinion of people can be obtained in order to see whether predicting the price is possible. Not only does Twitter offer real-time cryptocurrency information, it also provides a broad source of people's perspective towards the cryptocurrency. Since investors frequently share their feelings. According to behavioural economics, emotions and sentiment have a significant impact on individual behaviour and decision-making. *The major purpose of this study is to test whether Sentiment analysis on Twitter data is a useful tool to forecast price swings of an alternative cryptocurrency called Solana.* Given a large number of freely accessible data from Twitter containing the sentiment of cryptocurrency investors and users. To find out the relationship between Twitter's Sentiment and the Historical data a Vector autoregression model will be used. A Vector autoregression model is a statistical model used to capture the relationship between multiple variables that are time-sensitive.

2 RELATED WORK

2.1 *Cryptocurrency*

The term "cryptocurrency" can be coined as a virtual form of money. Cryptocurrency is a digital asset that functions as a medium of exchange with the use of encrypted transactions to make it secure. These encrypted transactions control the generation of new cryptocurrencies as well as the verification process of asset transfers. Cryptocurrencies are digital currencies that differ from traditional mediums of exchange in that they are based on the notion of decentralized control, as opposed to the reliance on banking services that occur in fiat money. Bitcoin is the first digital currency that led to the existence of cryptocurrencies.

The term "Bitcoin" got its first appearance in a publicly released paper by (Nakamoto, 2008). An electronic payment system based on cryptographic proof. Enabling transactions between parties without the involvement of a third party. Sellers would be protected from fraud by transactions that are computationally infeasible to reverse, and buyers would be protected by standard escrow techniques. One year later, the same author implemented the software for this peer-to-peer electronic cash system as open-source code (Nakamoto, 2009). The concept of Bitcoin is relatively simple to grasp. The most used example is to treat it as a puzzle (Nofer, Gomber, Hinz, & Schiereck, 2017; Tandon, Revankar, & Parihar, 2021). In the process of solving this puzzle, there is no central ledger. To carry out its operations, Bitcoin makes use of blockchain. Every block in the blockchain has a history of previous transactions. Every new transaction or service must be entered in the blocks of the blockchain. Adding the blocks to the blockchain. To give an idea, the minimum amount of functions in a single block is at least 2,500. Validation of each block is mandatory. This validation is done by crypto miners. In order to validate one block, each of the crypto miners has to answer to the last block. The person who solves the puzzle first is rewarded with twelve new Bitcoins. The procedure then begins all over again. Every ten minutes a new puzzle is released and it is made sure that it takes approximately ten minutes to solve the puzzle. The difficulty of the puzzle is adjusted based on how many computers are being used to "hunt" the Bitcoins (Houy, 2014).

2.2 *Solana*

These contributions led to a sparked wave of public interest, prompting others to create alternative cryptocurrencies based on the same principles but with a varying range of purposes. Such as Ethereum, Cardano, and Solana. The difference in these cryptocurrencies is that Bitcoin purely functions as a currency while the formerly called alternatives combine these principles to create applications like non-fungible tokens and ledger technologies on which companies can build new programs (Chohan, 2017).

Solana differs from the current standard of a blockchain infrastructure by making use of a procedure called "Proof of History". The core innovation that underlays Solana is a fundamental move in the structure of blockchain networks in regards to speed and capacity. While many people view crypto coins only as a virtual form of money, it is useful to think of crypto coins as a token that enables other applications on a platform. For example, Solana can power supply chain management contracts, managing electronic medical records, function as a peer-2-peer network for voting, governance systems, and more (Taskinsoy, 2019).

Moreover, Solana is also different from Bitcoin in the currency supply mechanism. Solana departed from Bitcoin by having opted for an unlimited supply of coins/tokens. It is expected to be inflationary in the first years after their release because the creation of new crypto coins is large in relation to the total stock that is already in circulation. However, in the long-run Solana's price will be more dependent on a strong demand for it to increase. That is when the absolute growth of Solana tokens imposed by the creation of new tokens will represent a modest share when the circulating supply is very large.

2.3 *Sentiment In Cryptocurrencies*

In contrast to normal stocks that are backed up with partial ownership of a business, cryptocurrencies are not backed by any underlying assets. The optimism and speculation of traders is what drives them forward (Wang, 2020). Traders believe they may sell the cryptocurrency to other people for a better price in the future, a concept known as the "greater fool theory". As a result, the driving force behind digital currencies' prices is speculation.

Subramaniam and Chakraborty (2020) states that cryptocurrency markets are mainly dominated by retail investors, individuals or non-professional investors who buy and sell securities through brokerage firms. Considering that individuals dominate the market and the earlier described driving force of digital currencies being speculation, the use of social media platforms such as Twitter to record the overall investor sentiment could be a viable option.

The price of currencies like the euro, as well as commodities, is determined by the interaction of supply and demand factors. In the context of cryptocurrencies, the impact of mining technology, which affects the supply side of the interaction while affecting the production part of the structure, have been investigated by Pizzol, Sacchi, Köhler, and Erjavec (2020). However, the mining of Bitcoin follows an algorithm that is publicly known and the level of demand is influenced by predictions about future price fluctuations as well as the underlying principles of the economy. For that reason, the efficient-market hypothesis may not fully capture the changes in Bitcoin prices, and short-term speculation investments or expectations should also be taken into account. These expectations could well be represented in common sentiment, bringing into question of how to gauge public sentiment and investigate its impact on price fluctuations of cryptocurrencies.

Matta, Lunesu, and Marchesi (2015) studied whether the increase of the price of Bitcoin was linked to the number of web searches or the number of tweets. They compared price patterns with tweet volume, in particular

positive tweets, as well as the Google Trend Data. By analyzing a total amount of 1.924.891 tweets marking them negative-positive, they found a correlation between the number of tweets and the Bitcoin price, as well as the Google Trend and Positive tweets. All have a statistical relationship with the price of the cryptocurrency.

Shah and Zhang (2014) used a Binary Autoregressive Tree Model (BART) to estimate the dynamics of the three largest cryptocurrencies. Predicting the direction of a change in the price rather than the value of change. In a ninety-day period of forecasting the dynamics of the cryptocurrencies, the predictive accuracy was around 64 per cent for Bitcoin, 62 per cent for Ethereum and 59 per cent for Ripple.

Karalevicius, Degrande, and De Weerdt (2018) used a percentage of positive and negative tweets to feed a recurrent neural network along with historical price data to predict the new price of Bitcoin. The accuracy for the classification of tweet's sentiment in a two class positive and negative was found to be 81,39 per cent and the overall accuracy of predicting the price using the recurrent neural network was 77,62 percent. Having a better accuracy than the mentioned paper by Shah and Zhang (2014), including text classification for the sentiment on top of the historical price data.

2.4 *Measuring Sentiment*

The present literature on using Twitter to predict fluctuations in alternative cryptocurrency prices with sentiment analysis is found to be constrained in a number of ways. Firstly, almost all of the prior research has mainly focused on the predictions of Bitcoin (Caviggioli, Lamberti, Landoni, & Meola, 2020; Pant, Neupane, Poudel, Pokhrel, & Lama, 2018; Stenqvist & Lönnö, 2017). Secondly, Apart from the promising work of Xie (2017), where the study obtained an 89 per cent accuracy of predicting the price changes in Bitcoin, there has been very little research to estimate cryptocurrency price returns using Sentiment analysis. In this study the sentiment was tested on Bitcointalk, a forum specifically made for cryptocurrency discussions. Thirdly, Karalevicius et al. (2018) shows that cryptocurrency investors exaggerate news leading to fluctuation patterns where the price follows the attitude of the Twitter users. Lastly, the aforementioned studies that did investigate alternative cryptocurrencies did not make use of a specialized lexicon that focuses on the use of emoji's specifically related to cryptocurrencies. Tweets sometimes only contain emoji's which the sentiment analysis would incorrectly score as neutral. Not making use of the importance an emoji can have.

This work offers a contribution by incorporating the use of specific cryptocurrency language with an alternative cryptocurrency that has not

been explored yet, namely "Solana". Many digital currencies have followed Bitcoin with having a maximum limit of the total supply put in circulation. Solana does not, which makes it even more interesting to find out whether this alternative cryptocurrency is affected by people's opinions. Within the crypto-asset market, the circulating supply metric is critical. It, together with the per-unit price of a crypto coin, allows investors to better comprehend the relative value of different assets. Considering this, the computational exploration of the sentiment of people's opinions could describe the price movements of Solana.

3 METHOD

3.1 *Data preprocessing and feature selection*

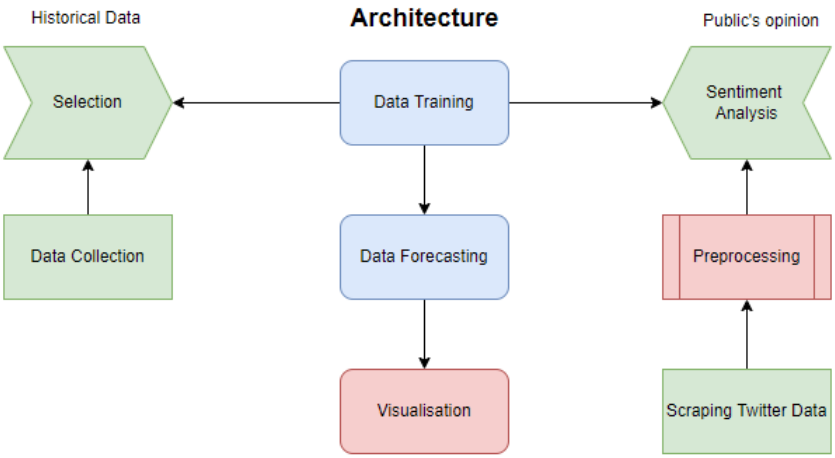

Figure 1: preprocessing and analyses steps

3.1.1 *Twitter Data*

The focus of this research is based on the Twitter sentiment regarding the cryptocurrency "Solana". Thus, all tweets containing the $sol were considered. For the scraping of these tweets, the python library Twint[1] was used, which stands for Twitter Intelligent Tool. This tool does not make use of any API which makes it fairly easy to work with. Twint also enables for selecting a language, a limit, and between specific dates. While scraping

[1] https://github.com/twintproject/twint

the tweets are automatically loaded into a CSV file. The parameters for scraping the tweets are that they are both in English and contain the $sol. In total there were 31065 tweets scraped in a time period between October 23 at 7 p.m and October 26 at 5 p.m. The following properties of the tweets are saved for the scraping Twitter data: time, date, and tweet. This is the process of scraping twitter data as seen in Figure 1.

After scraping the tweets have to be further pre-processed for them to generate meaningful sentiment scores. The sentiment analysis module, which will later be explained functions solemnly on English vocabulary. However, in the light of cryptocurrency people tend to use a form of slang to express their feelings. Examples of this are, "rocket", "moon", and "hodl" which in short means that the prices will go up. With that in mind, the tweets were first further pre-processed with the following steps: deleting all the hashtags and dollar sign words, but keeping the words themselves, since they contain some form of meaning. Scanning through all the words of the tweets and comparing them with the NLTK (Loper & Bird, 2002) corpus words as well as a customized set of words specific for cryptocurrency as described above. All the non-english words and/or words that are not in the customized set will be deleted. After this process, tweets that are "empty" are also discarded. At last, a preprocessing library in python for Twitter data is used to tokenize and clean the tweets. The purpose of tokenization is to protect sensitive data while saving its business use. It is the process of turning meaningful pieces of data such as price numbers into a random string of characters. For example, changing numbers to their word equivalents. The use of all these preprocessing steps resulted in a compressed data set of 19,216 tweets. Deleting a total number of 11,849 tweets. All of the above described steps are part of the preprocessing in Figure 1. Applying these techniques improves the quality of the data making it much more suitable for sentiment analysis. One example to describe the relevance of preprocessing is that tweets dominated by non-English words would lead to skewed neutral scores. While neutral tweets are of importance skewed data can be misleading and the analysis would not make the best use of the information present in the data set.

3.1.2 *Historical Price Data*

Cryptodatadownload[2] is a website that provides free historical cryptocurrency data in several time series. Binance, a cryptocurrency platform that is, according to coinmarketcap[3], the biggest, as well as the best provider for trading, was selected to get the best possible data. A CSV file containing

[2] found at: www.cryptodatadownload.com
[3] https://coinmarketcap.com/nl/rankings/exchanges/

the data can be obtained for free. After downloading the data, the date and time were set to be equal to the sentiment time series plus 12 hours for forecasting. The sentiment time series will be discussed later in this section. The closing price as well as the time and date were selected for modelling. The dates in which the price data is collected ranges from October 23 at 7 p.m to October 26 at 5 p.m. The data selected consists of 84 data points representing hourly closing prices of the cryptocurrency. This is part of the data selection in Figure 1. The predictive power of sentiment analysis in Twitter is detected to be best built between one to four days (Bollen et al., 2011; Zhang, Fuehres, & Gloor, 2011).

3.1.3 *Sentiment Analysis*

For the analysis of people's opinions the text analysis method called VADER (Valence Aware Dictionary for sEntiment Reasoning) by Hutto and Gilbert (2014) is used. This model is sensitive to both the polarity of a tweet as well as the intensity of people's emotions. The model maps lexical features to opinions known as sentiment scores. For example, the word 'sad' conveys a negative sentiment. Besides that, it is also able to understand the context of these words, so that "was not happy" can be seen as a negative statement. VADER is a rule-based sentiment analysis tool that is especially attuned to sentiments expressed in social media. Consequently, The model is also able to extract the meaning of emoji's. The aforementioned is one of the motivations why this sentiment analysis tool is great for social media. Furthermore, the model is able to understand the significance of the use of capitalization, such as "LOVE". The model calculates polarity scores of tweets, determining if the tweet expresses a positive, negative or neutral attitude. The polarity scores range between minus one and one. VADER computes a normalised weighted compound score indicating whether a tweet is positive, neutral or negative. The compound score metric calculates the sum of all the lexicon ratings which have been normalised. Positive sentiment has an compound score of larger than or equal to 0,05 and negative sentiment score of smaller than or equal to 0,05. Neutral being all the values in between. The polarity scores are then converted into time series of an hour. Taking the average of the tweet's sentiment per hour. Accounting for the effects of user influence is out of the scope of this research. This computation led to a data set of 72 hours containing the average sentiment per hour per day of over 19,000 tweets. Being the sentiment analysis step in Figure 1.

3.2 *Forecasting Models*

3.2.1 *Autoregressive Models*

The autoregressive (AR) model is a time series model that uses observations from previous time steps as input to a regression equation to forecast the value at the next step of the time. An autoregressive model thus predicts future behaviour based on past behaviour. The relationship between the present and past values are calculated by the correlation. The measurement is taken by calculating the relationship between the output variable and values at earlier time steps while using different lags. The stronger the correlation between the lagged variable and the output variable, the more weight the AR model puts on that variable when modelling. Given a multivariate time series, the Vector Auto Regressive Moving Average (VARMA) procedure approximates the models' parameters to create forecasts related to vector autoregressive moving-average processes. The AR and VARMA models are part of the data training and data forecasting part in Figure 1. By using multivariate time series opens the ability to include the sentiment and allow for external variables as well. By comparing the AR and VARMA models performance assumptions can be made about the predictability of Twitter's sentiment in forecasting the fluctuations of Solana's price. Training and testing the models for predicting the price fluctuations are identical in both models. An 70-30 split is partitioned in which 70 per cent of the data is reserved for training and 30 per cent is marked for testing. Alongside an additional 12 hours for the actual forecasting. After training the models, the actual predicting is compared to find out which one performs better.

3.3 *Evaluation*

3.3.1 *Linear Regression*

In order to find out whether time series models are indeed a good predictor for the price of a cryptocurrency linear regression is used to function as a baseline. It is the most basic regression model in machine learning. Linear regression is a linear approach for modelling the relationship between the dependent and independent variables. In prediction linear regression can be used to fit a predictive model to a data set consisting of response and explanatory variables.

3.3.2 *Augmented Dickey Fuller Test*

Considering that both our sentiment and Solana data are time-series data, it is important to check for stationarity in the data sets. To evaluate this, the Augmented Dickey-Fuller (ADF) test is used. An ADF tests the null hypothesis that a unit root is present in a sample of a time series. Unpredictable findings in time series analysis are caused by these unit roots. The ADF is a significance test, so the stationarity of the data is based on the test statistic. The ADF calculates the track of data and evaluates the degree to which a trend defines a time series. By having stationary data, the statistical properties do not depend on time. When exploring the historical price data, this was not the case. Differencing between consecutive observations helps stabilize the mean of a time series by removing changes in the level of a time series, and thus reducing trend and seasonality. The data used for the historical price was non-stationary. The Solana historical price is converted to stationary data for the time series models.

3.3.3 *Lag Order Autoregressive Models*

After a times series has been stationarized by differencing, the next step in fitting an autoregressive Model is to discover whether the AR terms are required to correct any residual autocorrelation in the time series after differencing. The systematic way to look at this is to plot both the autocorrelation function and the partial autocorrelation function differenced series. The autocorrelation presents how well the present value of the series is connected with the past values. The partial autocorrelation function finds the correlation of residuals, which remains after taking out the effects of earlier lags with the next lag value, therefore partial as already found variations are removed before finding the next correlation. The next lag can be modelled containing any hidden information in the residuals. So that this feature will be kept while modelling. For the autoregressive model the optimal lag order was to be found (10,0) (see appendix 7 Figure 2).

The VARMA model has a different procedure in which an integrated function can be used for selecting the order. The selection order automatically calculates the Akaike information criterion, Bayesian information criterion, and final prediction error for the estimated model. The selection order function then highlights the best possible lag order in which the former criterion is minimized. The asterix highlights the minimum which is the most suitable for the model. The ideal lag order for the VARMA model is (2,0) (see appendix 7 Table 3).

3.3.4 *Performance*

The square root of the variance of the residuals is the root mean square error (RMSE). It displays the model's absolute fit to the data, or how closely observed data points match the model's anticipated values. The root mean square error (RMSE) is an absolute measure of fit. The standard deviation of the unexplained variance can be viewed as the square root of the variance. The RMSE has the same units as the response variable, which is a helpful feature. A lower RMSE value indicates a better fit.

4 RESULTS

The results from the sentiment analysis indicated that in general, the public opinion about Solana is rather positive. The values range between 0.173 and 0.326 meaning that every hour the average sentiment opted to be positive. However, the variations between the hours say something about the negativity in the tweets as well. Lower hourly sentiments arise from the fact that there are more negative tweets. The central question however is whether sentiment scores have a predictive quality in forecasting the price fluctuations.

In total three models are being used to predict the price fluctuations in the cryptocurrency "Solana". The results for the root square mean error can be found in Table 1. For comparing the models to the linear regression the root mean square error is calculated of how good the model performs in predicting the values of the test set. While for comparing to autoregressive models the RMSE of the actual predicting is a better benchmark of performance. As shown in Table 1, the linear regression model does significantly worse in predicting the values ($RMSE = 10.005$) as both the autoregressive models ($RMSE = 2.490$ and $RMSE = 1.398$). One of the main reasons for this is that tweets do not have an immediate impact on the price. The linear regression model can not account for the lag relevant in time series models. To find out whether the hourly intervals of sentiment have a predictive power over Solana, a comparison is made by the autoregressive model and the Vector Auto-Regressive Moving Average model. While the former predicts values only making use of historical price

Table 1: Root mean square error results

Models	RMSE	Prediction RMSE
linear regression	10.005	
AR model	2.490	2.042
VARMA model	1.398	2.010

data, the latter does the same including sentiment scores. Looking at the actual predicting part of the models, the performance is almost identical, respectively $RMSE = 2.042$ and $RMSE = 2.010$. However, when looking at appendix 8 Figure 3 and appendix 8 Figure 4, which represent the AR model, the prediction lines clearly show that it tries to fit the data, while appendix 8 Figure 5 and appendix 8 Figure 6, which represent the VARMA model, the prediction lines are just flat. Showing that future predictions do not change over time.

A Granger causality test was done for determining whether the sentiment time series is useful in forecasting the historical price data. Measuring the ability to predict the future values of the historical price data using prior values of the sentiment time series. The Granger-Causality tests are checked for every number of lags useful for the model. The best number of lags opted to be 10, $p = 0,6548$ (see Appendix 7 Table 2). Showing that the acquired sentiment scores did not have a causal relationship with the price of Solana.

5 DISCUSSION

The goal of this study was to test whether sentiment analysis on Twitter data was a useful tool to predict price fluctuations of an alternative cryptocurrency called "Solana". As the results show time series analysis outperforms the linear regression benchmark by a lot. The reason the time series models outperformed the linear regression was due to the fact that there can not be an immediate effect of a tweet towards the price. Nevertheless, adding sentiment analysis on top of time series analysis to further increase the predictability did not enhance the model's performance. *Concluding that sentiment analysis alone in combination with historical price data is not a useful tool to predict price fluctuations of Solana*. An interesting finding was that the autoregressive model that predicted the price values by using previous values performed very similar to the model with including the sentiment. The future predictions including the sentiment were pretty much the same as all the predictions. Making the model unable to generalize. This happens when the stationary scores do not have strong seasonality. Making it difficult for the VARMA model to predict future data points, eventuating in that it takes the average of previous values and predict it as future values. Therefore producing a flat line. As we can see from Figure 5 and 6. An explanation for this is Solana's volatility, since it is still at a very nascent stage compared to well-established investment tools and currencies (Chu, Chan, Nadarajah, & Osterrieder, 2017).

While the analysis of Matta et al. (2015) conclude that sentiment analysis contributes to be seen as a predictor, this research indicates that for Solana

this is not the case. A major difference between this research and the research of Matta et al. (2015) and Karalevicius et al. (2018) is the number of tweets. The first made use of over 1,900,000 tweets regarding Bitcoin and the latter of 190,000 tweets both being notably higher than the mere 19,000 tweets of this research. A larger data set increases the probability that it contains useful information, which leads to lower estimation variance and thus better predictive performance.

The data set consisted of over 19,000 tweets containing the dollar sign "SOL" of which one of the issues is the contextual relevance due to the fact that these search terms could be of no use. For example, tweets related to Solana could use the hashtag to gain some attention, while not actually talking about Solana. Resulting in tweets that might contain the word of interest, yet the information in the tweet is not relevant for this Sentiment Analysis.

While sentiment analysis should surely not be discarded as a predictor, according to the results of this study it can be concluded that it should be an addition to and not the predictor for price changes. Other additions that could increase the models' performance are weighing the tweets with the follower count, finding a way to avoid tweets containing hashtags but are of no relevance to the topic, and the relevance of retweets.

This work offered a contribution to the limited amount of research done in cryptocurrency by testing whether sentiment analysis, is a useful tool to increase a models performance in predicting Solana's price fluctuations. An alternative cryptocurrency on which no research has been done. This by making use of the VADER algorithm and VARMA model to forecast these price changes. Sentiment analysis is a procedure that requires a high rate of analysis depth to be able to make accurate findings. It is also worth noting that tweets are relatively noisy as data, thus making inferences from them should be done with caution.

6 CONCLUSION

This study investigates whether solely the sentiment of tweets regarding an alternative cryptocurrency is a good predictor for the fluctuations in price. Using the Valence Aware Dictionary for Sentiment Reasoning for sentiment analysis along with the natural language toolkit for preprocessing to give sentiment scores to tweets. While previous studies have mainly focused on cryptocurrencies with an already acquired market position, this study focuses on a rapidly growing cryptocurrency called Solana.

The predictive performance analysis is carried out over a period of 74 hours in which 31,000 tweets were collected to predict the price changes in Solana. Next to this the historical price data for the same period plus

twelve hours for forecasting was collected. The basic time series model made use only of the historical price data to predict future prices, while the multivariate one included the sentiment scores. In this study, the multivariate time series model (RMSE = 2,042) did not perform better than the univariate one (RMSE = 2,010).

In order to further improve the accuracy of Sentiment analysis, additional research can be done in obtaining a better data set containing tweets solely about Solana. Discarding tweets that are either full of hashtags or have a maximum amount of cryptocurrencies as hashtags to diminish irrelevant tweets.

Unlike most protocols in cryptocurrency blockchain that run with the Proof of Stake or Proof of Work mechanism, Solana uses Proof of History, a new cryptographic procedure that amplifies scalability while maintaining security. An important feature is that the Proof of History mechanism reduces the time between transactions with much lower fees in comparison to other alternative cryptocurrencies. On the strength of this protocol, Solana's debut attracted high-profile companies in the blockchain. Over 230 companies are currently included in the ecosystem of Solana. A lot of these partnerships are established by the explosive growth of non-fungible tokens on the Solana blockchain. A suggestion for future researchers is to be looking for tweets that concern partnerships between companies and Solana. The Solana network is designed to be a high-performance blockchain. These design decisions mean that the network is very energy efficient in opposite to other blockchain networks. Transactions are the fundamental building blocks of blockchain and an increasing number of companies are now investing or making use of this still very new technology. Partnership announcements could have an impact on how confident people are in a certain cryptocurrency.

REFERENCES

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1–8.

Caviggioli, F., Lamberti, L., Landoni, P., & Meola, P. (2020). Technology adoption news and corporate reputation: sentiment analysis about the introduction of bitcoin. *Journal of Product & Brand Management*.

Chohan, U. W. (2017). Assessing the differences in bitcoin & other cryptocurrency legality across national jurisdictions. *Available at SSRN 3042248*.

Chu, J., Chan, S., Nadarajah, S., & Osterrieder, J. (2017). Garch modelling of cryptocurrencies. *Journal of Risk and Financial Management*, *10*(4), 17.

Houy, N. (2014). The bitcoin mining game. *Available at SSRN 2407834*.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

Karalevicius, V., Degrande, N., & De Weerdt, J. (2018). Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance*.

Kemdal, A. B., & Montgomery, H. (2002). Perspectives and emotions in personal decision making. In *Decision making* (pp. 86–103). Routledge.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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). 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.

Nakamoto, S. (2009). Bitcoin v0. 1 released. *The Mail Archive*, *9*.

Nofer, M., Gomber, P., Hinz, O., & Schiereck, D. (2017). Blockchain. *Business & Information Systems Engineering*, *59*(3), 183–187.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, *6*(3), 144–160.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (scopes)* (pp. 1345–1350).

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018). 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).

Pizzol, M., Sacchi, R., Köhler, S., & Erjavec, A. A. (2020). Non-linearity in the life cycle assessment of scalable and emerging technologies. *Frontiers in Sustainability*, *1*, 13.

Shah, D., & Zhang, K. (2014). Bayesian regression and bitcoin. In *2014 52nd annual allerton conference on communication, control, and computing (allerton)* (pp. 409–414).

Stenqvist, E., & Lönnö, J. (2017). *Predicting bitcoin price fluctuation with twitter sentiment analysis.*

Subramaniam, S., & Chakraborty, M. (2020). Investor attention and cryptocurrency returns: Evidence from quantile causality approach. *Journal of Behavioral Finance*, *21*(1), 103–115.

Tandon, C., Revankar, S., & Parihar, S. S. (2021). How can we predict the impact of the social media messages on the value of cryptocurrency? insights from big data analytics. *International Journal of Information Management Data Insights*, *1*(2), 100035.

Taskinsoy, J. (2019). Blockchain: a misunderstood digital revolution. things you need to know about blockchain. *Things You Need to Know about Blockchain (October 8, 2019)*.

Wang, M. (2020). Bitcoin and its impact on the economy. *arXiv preprint arXiv:2010.01337*.

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, *37*(2), e12493.

Xie, P. (2017). *Predicting digital currency market with social data: Implications of network structure and incentive hierarchy* (Unpublished doctoral dissertation). Georgia Institute of Technology.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, *26*, 55–62.

7 APPENDIX A

Table 2: Granger-Causality tests

number of lags	p-value	f test
1	0.9579	0.0028
2	0.9955	0.0045
3	0.9360	0.1395
4	0.7559	0.4722
5	0.7799	0.4932
6	0.8267	0.4697
7	0.7211	0.6394
8	0.8504	0.4992
9	0.7176	0.6856
10	0.6548	0.7716

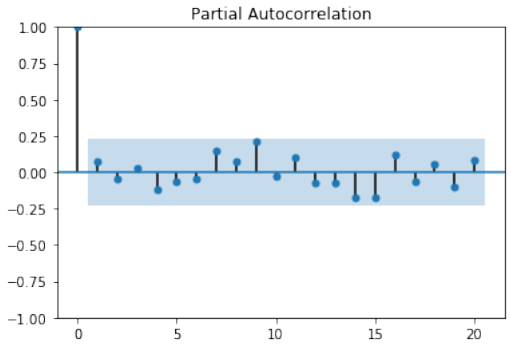

Figure 2: Partial autocorrelation plot

Table 3: Selection order * highlights the minimum

	AIC	BIC	FPE	HQIC
0	-3.779	-3.699	0.02284	-3.749
1	-4.167	-3.926	0.01551	-4.077
2	-4.408*	-4.006*	0.01220*	-4,258*
3	-4.291	-3.729	0.01376	-4.082
4	-4.177	-3.454	0.001552	-3.907
5	-4.145	-3.262	0.01616	-3.816
6	-4.127	-3.083	0.01669	-3.738
7	-4.042	-2.838	0.01851	-3.593
8	-4.204	-2.839	0.01612	-3.672
9	-4.240	-2.715	0.01612	-3.672
10	-4.264	2.578	0.01645	-3.636

8 APPENDIX B

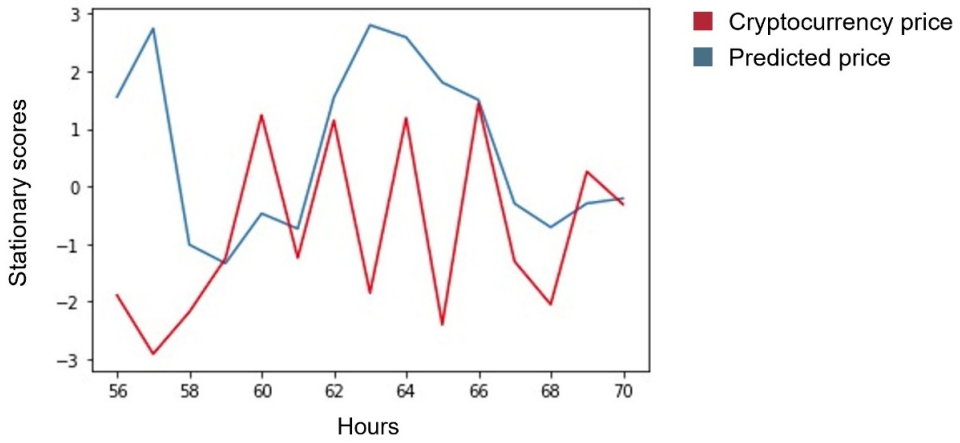

Figure 3: AR model RMSE

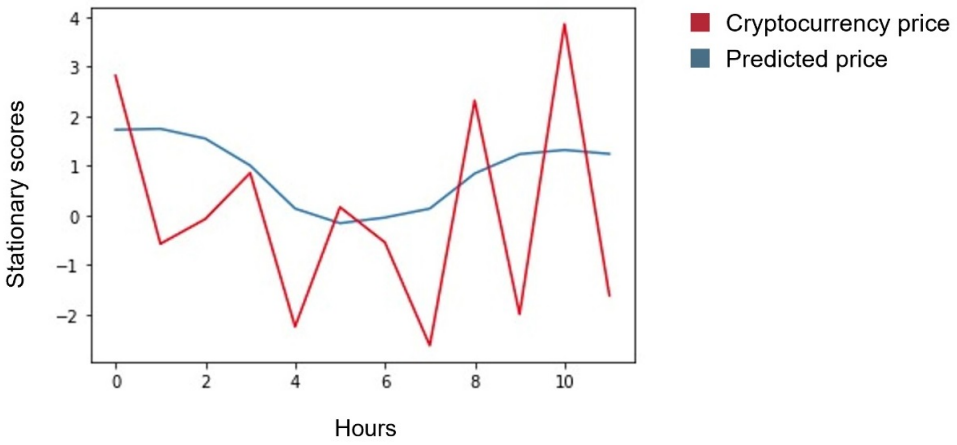

Figure 4: VARMA model RMSE

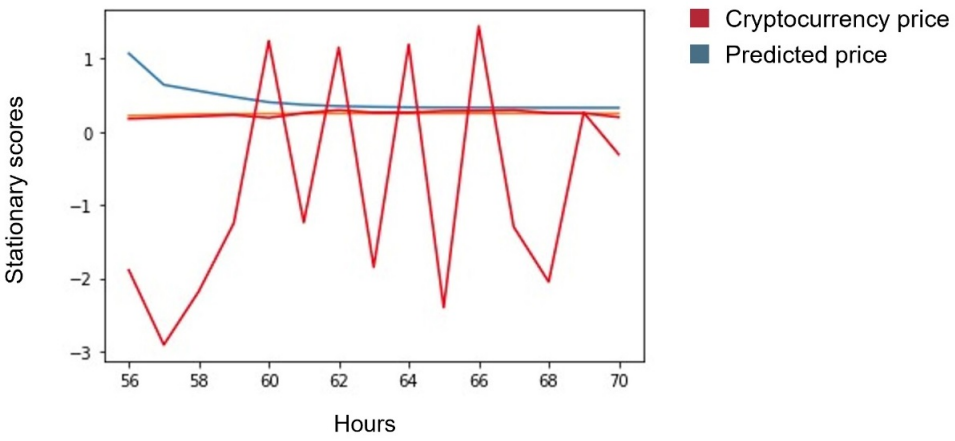

Figure 5: AR model prediction RSME

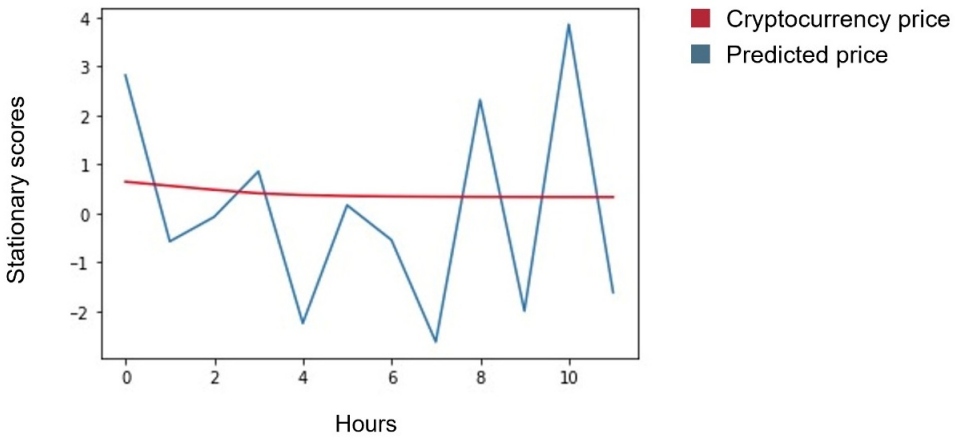

Figure 6: VARMA prediction model RMSE