Title	Predicting cryptocurrency price movement and magnitude using alternative data
Sub Title	
Author	Rodenberg, Braedon B(Hayashi, Takaki) 林, 高樹
Publisher	慶應義塾大学大学院経営管理研究科
Publication year	2022
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2022年度経営学 第4029号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003001-00002022-4029

慶應義塾大学大学院経営管理研究科修士課程

学位論文（2022年度）

論文題名

Predicting Cryptocurrency Price Movement and Magnitude Using Alternative Data

主　査	林　高樹
副　査	山尾　佐智子
副　査	武田　史子
副　査	

氏　名	ローデンバーグ　ブレイデン B

論 文 要 旨

所属ゼミ	林 高樹 研究会	氏名	ローデンバーグ　ブレイデン　B

（論文題名）

Predicting Cryptocurrency Price Movement and Magnitude Using Alternative Data

（内容の要旨）

Cryptocurrencies have gained popularity in recent years due to their decentralized nature, anonymity, low fees, and accessibility, but they also suffer from high volatility, lack of regulation, limited acceptance, and complexity. With that popularity has come research aiming to analyze correlations and predict price movement trends using alternative data sources (information not typically used in financial analysis), such as Twitter and Reddit, due to Tweets from public figures and Reddit community movement swaying cryptocurrency prices.

Recently, Critien et al. (2022) attempted to predict not only the trend in Bitcoin price movement but the magnitude, how much that increase or decrease would be, using cryptocurrency price data and Twitter as an alternative data source. Due to existing studies reporting a correlation between Bitcoin price movement and Reddit communities (Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S., 2019), we conjecture that data collected from Reddit's longer-form discussions and exchange of information within cryptocurrency-conscientious communities may provide additional predictive power for Bitcoin prices. Based on Critien et al. (2022)’s approach as a benchmark, this research attempts to increase magnitude prediction accuracy scores by using Reddit data (volumes and sentiment scores) in addition to Twitter and cryptocurrency price data as explanatory variables.

In line with Critien et al. (2022)’s research, we use CNN, LSTM, and BiLSTM neural network models on time series data from 30 August 2018 to 23 November 2019, gathered from openly available datasets on Kaggle, and incorporate time lag as an additional parameter and sentiment scores as an explanatory variable.

Our results show that adding Reddit data leads to a statistically significant increase in mean and max accuracies for certain time lags in general. However, there is not a statistically significant increase in accuracy when comparing the highest-performing models of ours and Critien et al. (2022)’s.

Table of Contents

1. Research Background

Cryptocurrencies, such as Bitcoin, have gained popularity since their creation in 2008 due mainly to decentralization, anonymity, low fees, and accessibility. They are not controlled or regulated by a central authority, all transactions can be anonymous, transaction fees can be lower than traditional methods of sending money, and anyone with an internet connection can access them. However, along with these benefits come high volatility in price, a lack of regulation – making them susceptible to fraud and scams, limited acceptance, and systemic complexity involving blockchain technology that is often hard for a typical user to understand.

With significant price increases seen in a relatively short amount of time (Bitcoin increased roughly 11,800% in value - $5,165 to $61,283 per coin - from March 13, 2020, to March 12, 2021; Dogecoin increased approximately 24,600% in value - $0.0026 to $0.64 per coin – from May 8, 2020, to May 7, 2021), cryptocurrencies have gained attention by corporations and financial institutions. Tesla revealed in early February 2021 that they purchased $1.5 billion worth of Bitcoin and would start accepting the cryptocurrency as payment (Stevekovach, 2021), after which the price nearly doubled in the space of a month.

Figure 1: Price of Bitcoin in USD (Source: Google Finance)

Retail investors have also become interested in cryptocurrency, increasing their total possession of Bitcoin's circulating supply from 12% in early 2020 to 17% as of December 2022 (Throuvalas, A., 2022). With retail investors gaining more skin in an unregulated market, there are opportunities for retail investors to collectively influence or be influenced by others through social media, causing artificial bubbles and crashes.

Twitter and Reddit, in particular, are two well-known social media platforms that have been said to influence cryptocurrency prices. The two platforms differ in their user demographics, and the types of content typically shared on each platform. While Twitter is primarily utilized for sharing timely updates and news in a concise format, Reddit is geared towards longer-form discussions and the exchange of information. This divergence in content leads to a diverse user base on Twitter, while Reddit concentrates more on communities and interests.

There is research and speculation about the effect Elon Musk and other social influencers' Twitter profiles and tweets have on cryptocurrency price volatility with their reach to the masses (Zaman, S., Yaqub, U. and Saleem, T., 2022), as well as how movement in specific Reddit discussion groups ("Subreddits") is correlated with cryptocurrency price movement (Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S., 2019) or been used to manipulate markets – creating artificial bubbles and crashes artificially (Sawhney et al., 2022). Data from these sources and other social media and news outlets (also known as "alternative data") have been combined with price movements to predict cryptocurrency price trends using machine learning models, as seen in Figure 2.

	Predict Movement/ Correlation/Bubbles		Predict Price
Financial Data			Marne et al (2021)
Financial Data & Alternative Data	Tandon et al (2021) Bitcoin/Dogecoin | **Twitter**	H Alejandro (2021) Bitcoin | **Twitter/Reddit**	Tandon et al (2021) Bitcoin | **Twitter**
	Lamon et al (2017) Bitcoin/Litecoin/Ethereum | **News/Social Media**	Wooley et al (2019) Bitcoin/Ethereum | **Reddit**	M Critien et al (2022) Bitcoin | **Twitter**
	Abraham et al (2018) Bitcoin/Ethereum | **Twitter/Google Trends**	C Phillips et al (2017) Bitcoin | **Twitter/Reddit**	Bremmer (2018) Ethereum/Litecoin/ Ripple | **Reddit**
	Gurrib et al (2021) Bitcoin | **News**	C Phillips et al (2017) Bitcoin | **Twitter/Google Trends/Reddit/ Wikipedia**	
	University of Warick Bitcoin | **News/Reddit**		

Figure 2: Cryptocurrency Movement and Magnitude Prediction Research Map

As cryptocurrency is an unregulated market, there are no restrictions on day trading or investment amounts. A machine learning model able to predict cryptocurrency price movement would allow a financial institution or retail investor to maintain a continuous flow of profit by buying and selling at any moment the model predicts an upward or downward trend. It would be even more practical for the model to be able to predict the magnitude of movement, or how much the value will go up or down, as it would allow investors to determine whether a bubble or crash is coming, as well as how much cryptocurrency to buy or sell at one time to maintain overall target profit levels.

Most alternative and price data research focuses on calculating correlation or predicting trends. Still, there is gaining interest in predicting the magnitude of movement, as seen by research published in May 2022, using Twitter and price data to predict bitcoin price movement trends and magnitude (Critien et al., 2022).

2. Research Purpose

The purpose of this research is to investigate whether Reddit data can be combined with Twitter and financial data (price over time, transaction history, etc.) to better predict cryptocurrency price movement and magnitude than the most recent research using only Twitter as an alternative data source (Critien et al., 2022), in hopes of contributing to future research focusing on price movement magnitude prediction. Critien et al. (2022)'s approach was used as a baseline for our research as it is the only known research to focus on the magnitude prediction of Bitcoin using Twitter and financial data. Models and data used and research results were openly published by the authors, making it easier to research the impact of adding Reddit data as an explanatory variable.

In a practical sense, improving the accuracy of cryptocurrency price movement and magnitude prediction models is important because it can be applied to a daily buy/hold/sell recommendation engine that automatically maximizes day-trading profits. A model using both price movement and magnitude, like the one in our research, would be more informative for investors than a recommendation engine that only shows price movement, as some investors would hold investments if they saw only a 1% decrease forecast for the next day, but would most likely sell if they saw a 30% decrease the next day; in the case of a price movement only recommendation engine, both would only show a generic price decrease the next day.

Furthermore, the higher the accuracy of the model, the higher the return over time (ROT) on investment when applied to a Buy/Hold/Sell recommendation engine; the University of Warwick showed a 287.9% ROT over seven months when using a model that predicted future price movement (no magnitude predictions) with 63.5% using news and Reddit data (University of Warwick).

	Predict Movement/ Correlation/Bubbles		Predict Price
Financial Data			Marne et al (2021)
Financial Data & Alternative Data	Tandon et al (2021) Bitcoin/Dogecoin | **Twitter**	H Alejandro (2021) Bitcoin | **Twitter/Reddit**	Tandon et al (2021) Bitcoin | **Twitter**
	Lamon et al (2017) Bitcoin/Litecoin/Ethereum | **News/Social Media**	Wooley et al (2019) Bitcoin/Ethereum | **Reddit**	M Critien et al (2022) Bitcoin | **Twitter**
	Abraham et al (2018) Bitcoin/Ethereum | **Twitter/Google Trends**	C Phillips et al (2017) Bitcoin | **Twitter/Reddit**	Bremmer (2018) Ethereum/Litecoin/ Ripple | **Reddit**
	Gurrib et al (2021) Bitcoin | **News**	C Phillips et al (2017) Bitcoin | **Twitter/Google Trends/Reddit/ Wikipedia**	TARGET
	University of Warick Bitcoin | **News/Reddit**		

Figure 3: Cryptocurrency Movement and Magnitude Prediction Research Map; Research Target Included

3. Novelty of Research

As mentioned earlier, the existence of correlations between Bitcoin price movement and Reddit communities has been reported in the literature (Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S., 2019), which naturally leads to conjecture about the existence of unique information contained in Reddit's longer-form discussions and exchange of information within cryptocurrency-conscientious communities. Nevertheless, utilizing Reddit data in combination with Twitter data to improve Bitcoin price movement magnitude prediction accuracy has yet to be attempted in the literature, so far as we are aware.

4. Research Method

This empirical research uses machine learning models to predict price movement direction and magnitude. The mean prediction accuracy and variance results of the Twitter dataset and machine learning models used in Critien et al. (2022) are used and compared with the same machine learning model results using Twitter and Reddit data combined to determine whether adding Reddit data as an explanatory variable can better predict future prices or reduce the prediction variance. Figure 4 shows the research process.

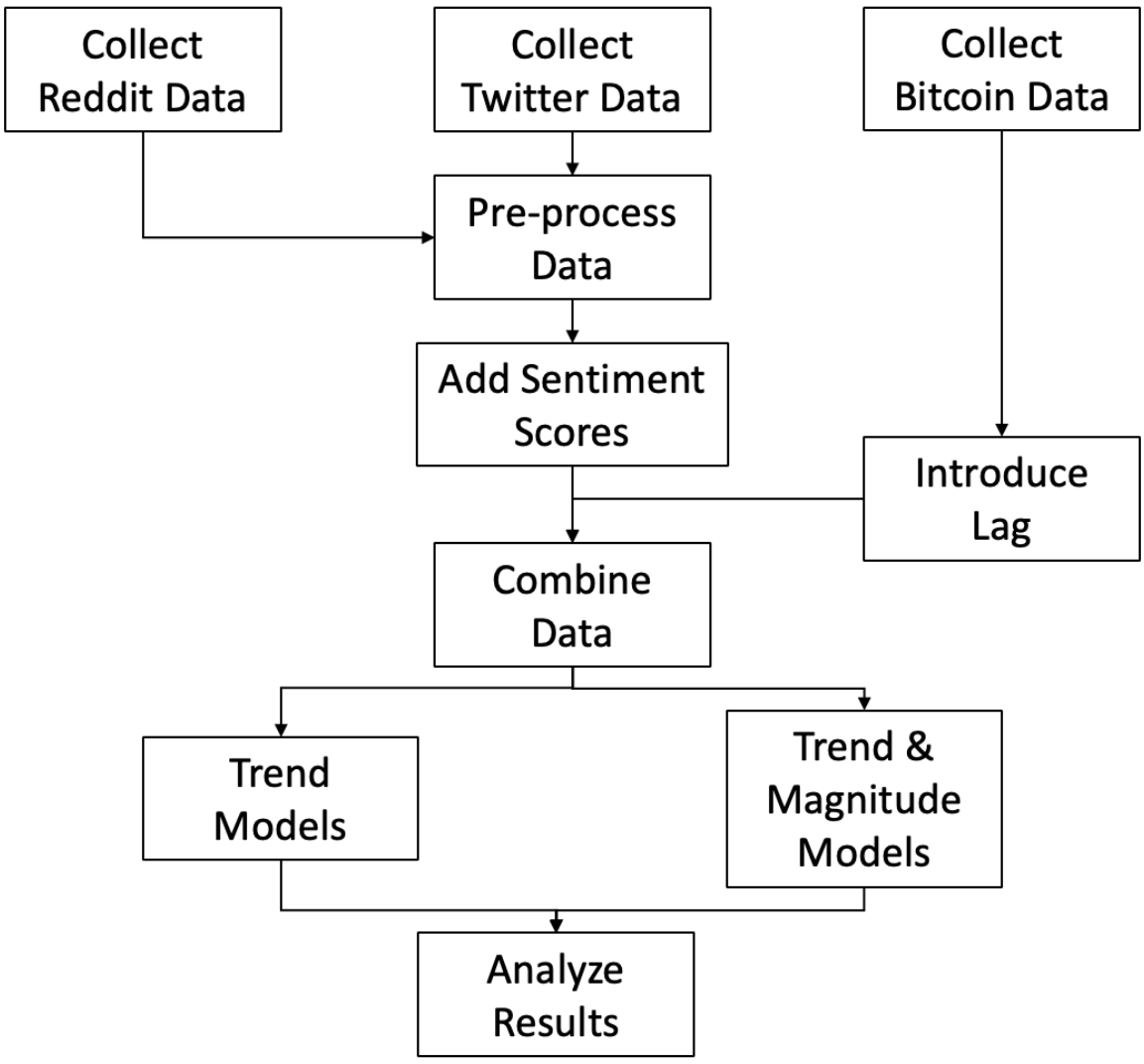


Figure 4: Research Method

Data Collection

Reddit data is collected from the "Reddit Comments Containing "Bitcoin" 2009 to 2019*"* dataset openly available on Kaggle (Faneli, 2021). Kaggle is a platform for data science competitions that allows users to find and publish datasets that cover a wide range of topics and industries. This dataset includes over 4 million Reddit comments containing the word "bitcoin", the date and time they were posted, the author, the subreddit in which they were published, how many up or downvotes the comment received, and the comment text. This research used all comments posted between 30 August 2018 and 23 November 2019, the same period used in Critien et al. (2022).

Twitter and Bitcoin data is collected from the same open Kaggle datasets as Critien et al. (2022), "Bitcoin tweets – 16M tweets" (Alex, 2019) and "Bitcoin Historical Data" (Zielak, 2022). The Twitter dataset includes any tweet containing "bitcoin" or "BTC" between 1 January 2016 and 29 March 2019, the poster's username, the date and time it was posted, how many likes, replies, and retweets it received, as well as the tweet text. The Bitcoin dataset includes the Bitcoin open, high, low, and close price and the volume of currency traded in 1-minute intervals between 1 January 2012 and 31 March 2021.

Data Pre-processing

Following Critien et al. (2022), collected alternative data (Twitter and Reddit) is pre-processed by first removing duplicates and non-English tweets and comments (this research focused on English alternative data as English is the primary language used among cryptocurrency investors). To prepare the data for sentiment analysis, URLs, hashtags, mentions, and punctuation are removed from the tweet/comment bodies, the bodies are lemmatized, and any row containing less than four words is removed.

Sentiment Scores

Critien et al. (2022) assigns sentiment scores to tweets to use public sentiment as an explanatory variable in the machine learning models, a method widely used in similar research (Abraham et al., 2018; Kraaijeveld and De Smedt, 2020; Mohapatra et al., 2020; Serafini et al., 2020; Valencia et al., 2019). The VADER-Sentiment-Analysis package, "a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media" (Hutto), is used to polarize tweets and comments with a sentiment score. VADER assigns negative, positive, neutral, and compound polarity scores to each comment/tweet body. The compound score is a sum of the individual sentiment scores normalized to fall within the range of -1 to 1. Sentiment scores are collected for both Twitter and Reddit data. Still, only the positive and negative polarity scores are included in the training and evaluation data sets following Critien et al. (2022).

Combining Data

Twitter and Reddit data are individually grouped by day for this research, taking the sum of tweets/comments and the average positive and negative polarity scores during the day period; the two datasets are then joined on the day. Once the Twitter and Reddit datasets are merged, the closing Bitcoin price for the day was added to the dataset.

Lag Introduction

Critien et al. (2022) considers the lag between tweet and price change when determining the optimal lag that shows the highest correlation for predicting price movement and magnitude, such as 1, 3, or 7 days. This-treatment considers how long a tweet or comment on social media takes to impact the cryptocurrency price.

Accordingly, three datasets are created from the **Combining Data** step for analysis – a 1-, 3-, and 7-day price lag dataset – each containing the same Twitter and Reddit data grouped by

date, but for example, the Bitcoin close price from the next day used instead of today's Bitcoin close price for the 1-day lag dataset.

Explanatory Variables

The explanatory variables used in the model are Bitcoin close price ("Close"), Twitter positive sentiment ("t_pos_pol"), Twitter negative sentiment ("t_neg_pol"), Tweet volume ("tweet_vol"), Reddit positive sentiment ("r_pos_pol"), Reddit negative sentiment ("r_neg_pol"), and Reddit comment volume ("reddit_vol").

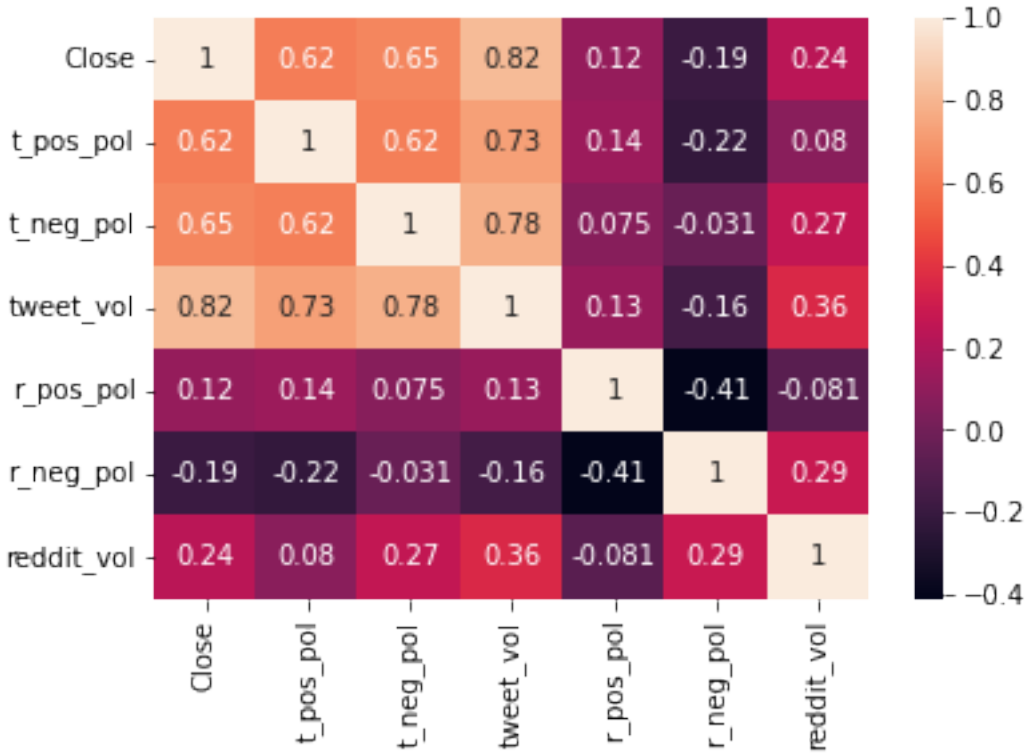


Figure 5: Explanatory Variable Correlation Matrix

It is worth mentioning that Tweet volume and Twitter sentiment are more highly correlated to the Bitcoin close price than other variables in Figure 5. Including multiple highly correlated variables in a model can cause multicollinearity, resulting in unstable and inconsistent coefficients and making it difficult to interpret the results. To address this issue, one can remove or combine correlated variables, use a different model less sensitive to

multicollinearity, or check for correlations before building the model. However, this research follows the same procedures as Critien et al. (2022) and does not remove highly correlated variables.

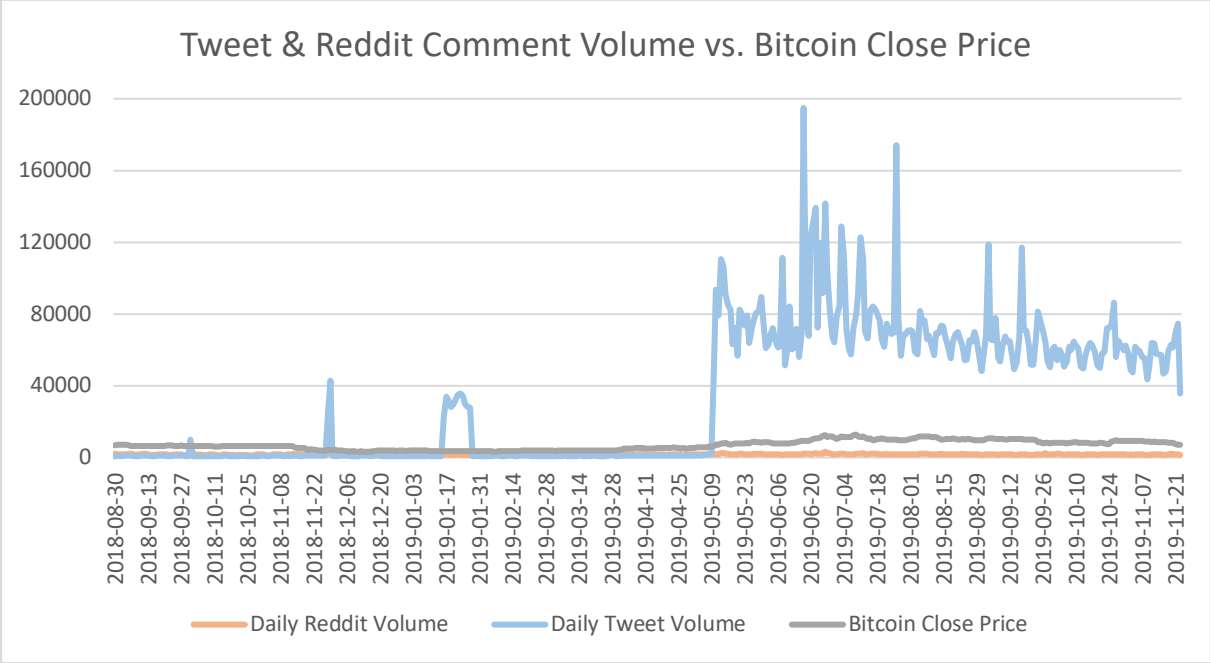


Figure 6: Tweet and Reddit Comment Volume vs. Bitcoin Close Price

During the selected time interval, Tweet volume increased significantly, followed by an increasing trend in Bitcoin close price. Although Bitcoin trade volume is not included in the model as an explanatory variable, Tweet volume follows a similar trend over time with Bitcoin trade volume, as shown in Figure 7.

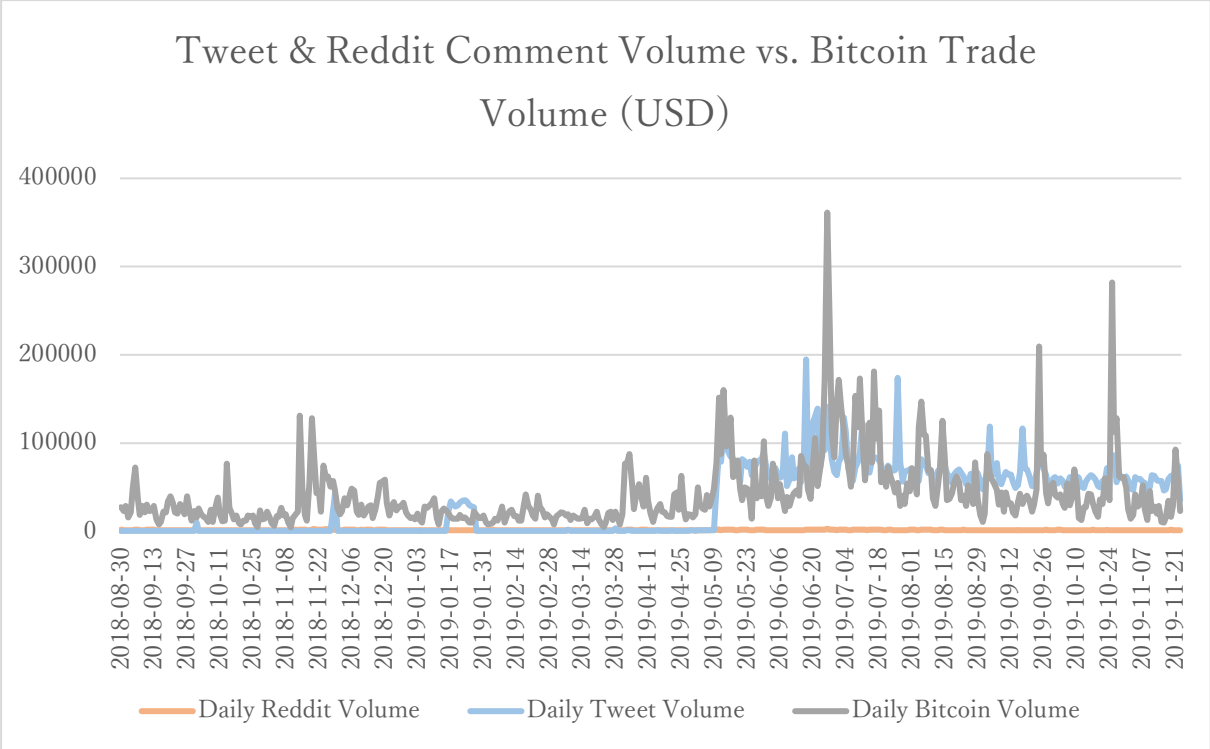


Figure 7: Tweet and Reddit Comment Volume vs. Bitcoin Trade Volume (USD)

Also worthy to mention is the change in Twitter sentiment over time, whereas Reddit stayed primarily consistent throughout the selected period, as seen in Figure 8 and Figure 9.

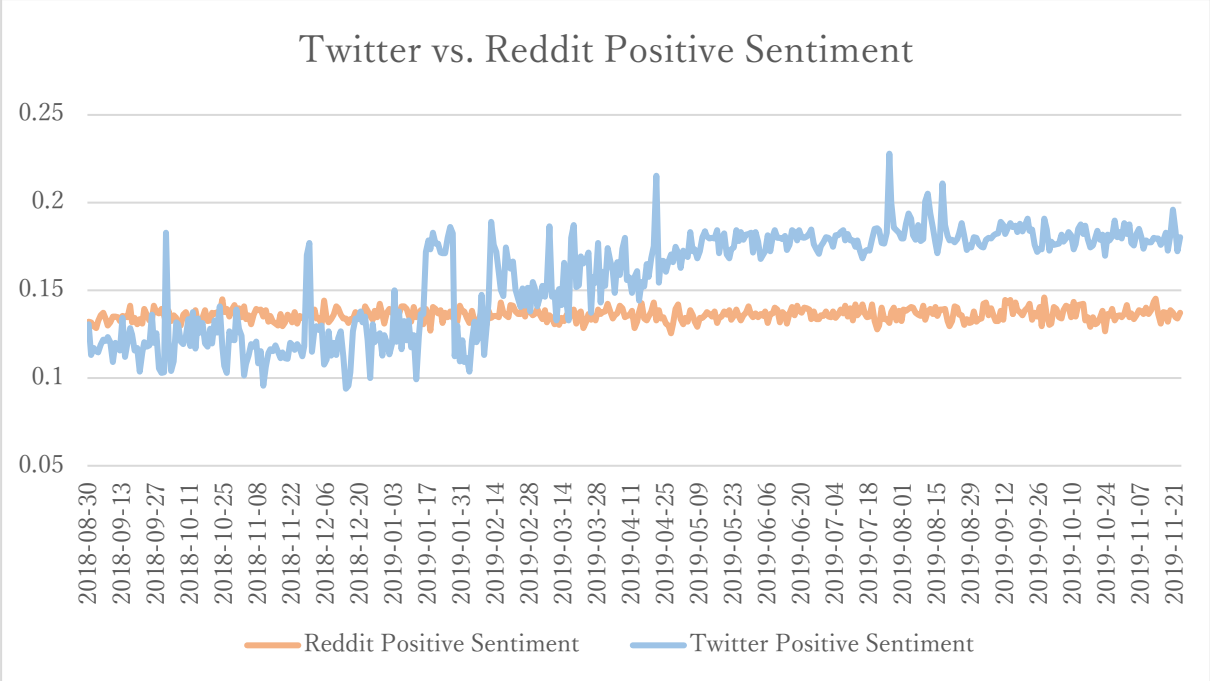


Figure 8: Twitter vs. Reddit Positive Sentiment

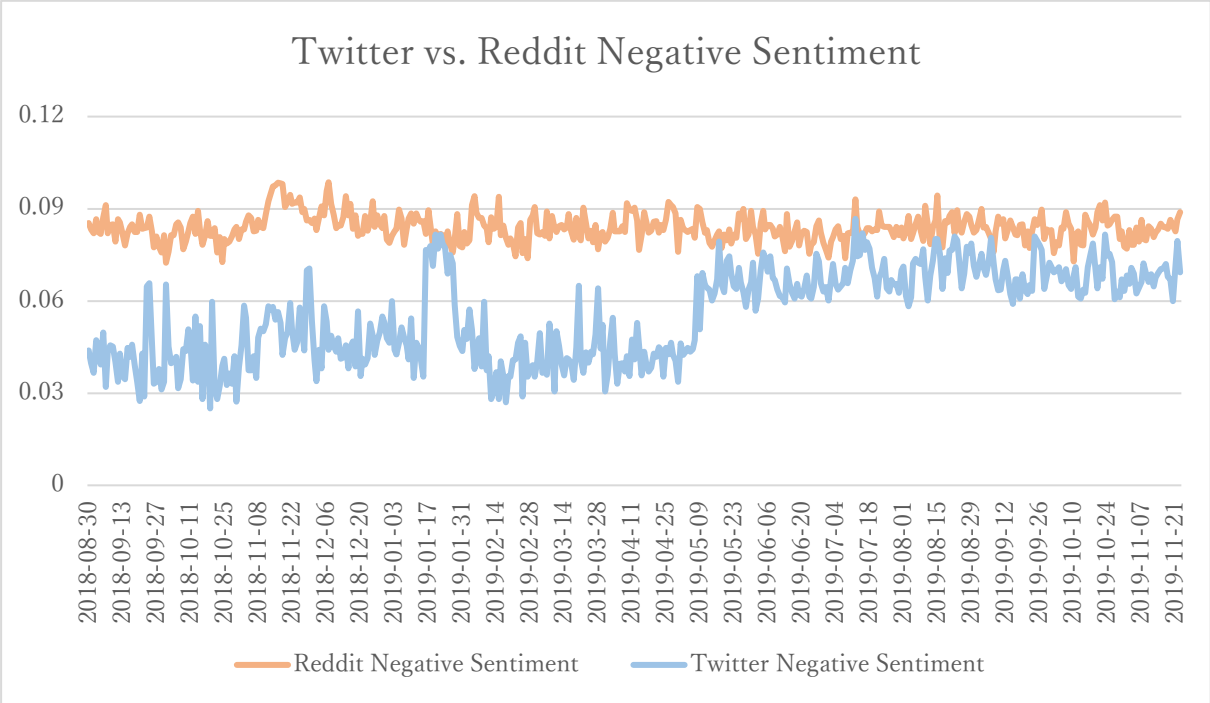


Figure 9: Twitter vs. Reddit Negative Sentiment

As seen in Figures 10 and Figure 11, there are two clusters of sentiment scores which can be divided as before May 2019 and after May 2019, when Twitter's price started to increase above $8,000 per coin, a new high record at the time. No significant news event or Tweet from a famous person could be directly tied to the increase. Still, it was speculated that "whale" traders (a person or group that trades large amounts of financial instruments, typically on financial markets. These trades can significantly impact market prices or volumes due to their size. Whale traders are often professional investors or financial institutions with the resources and capital to place large trades) speculating about future price increases were the cause (Reid, 2019).

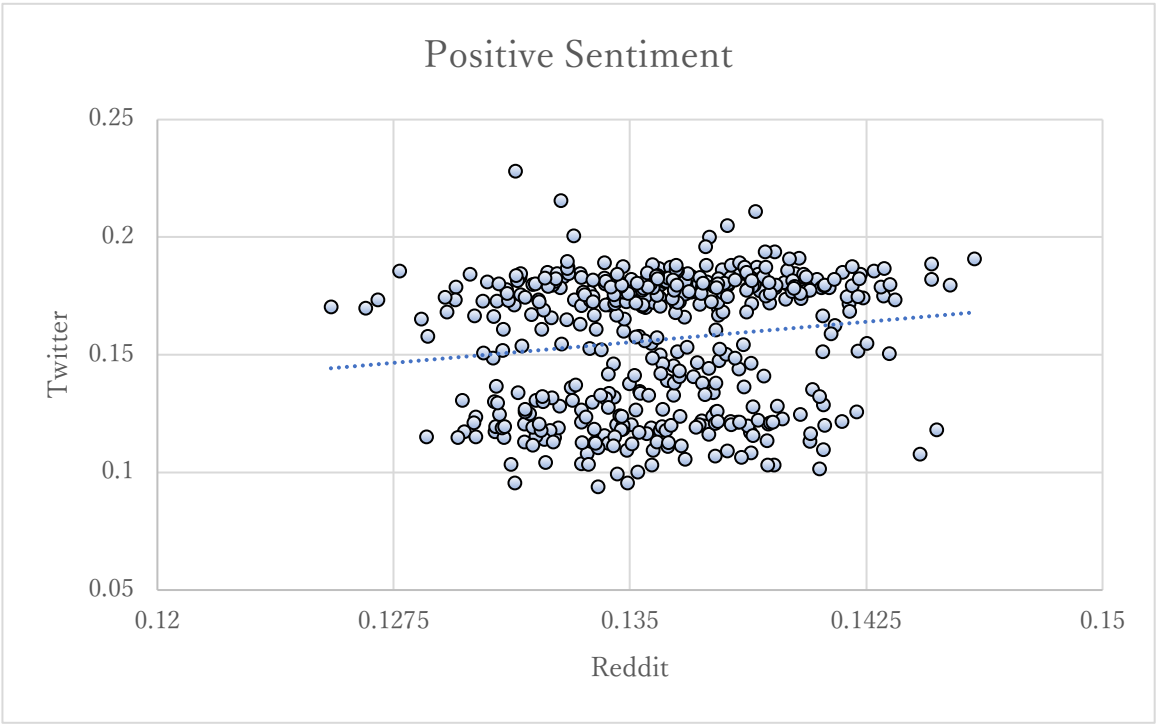


Figure 10: Twitter vs. Reddit Positive Sentiment Correlation (r = 0.143)

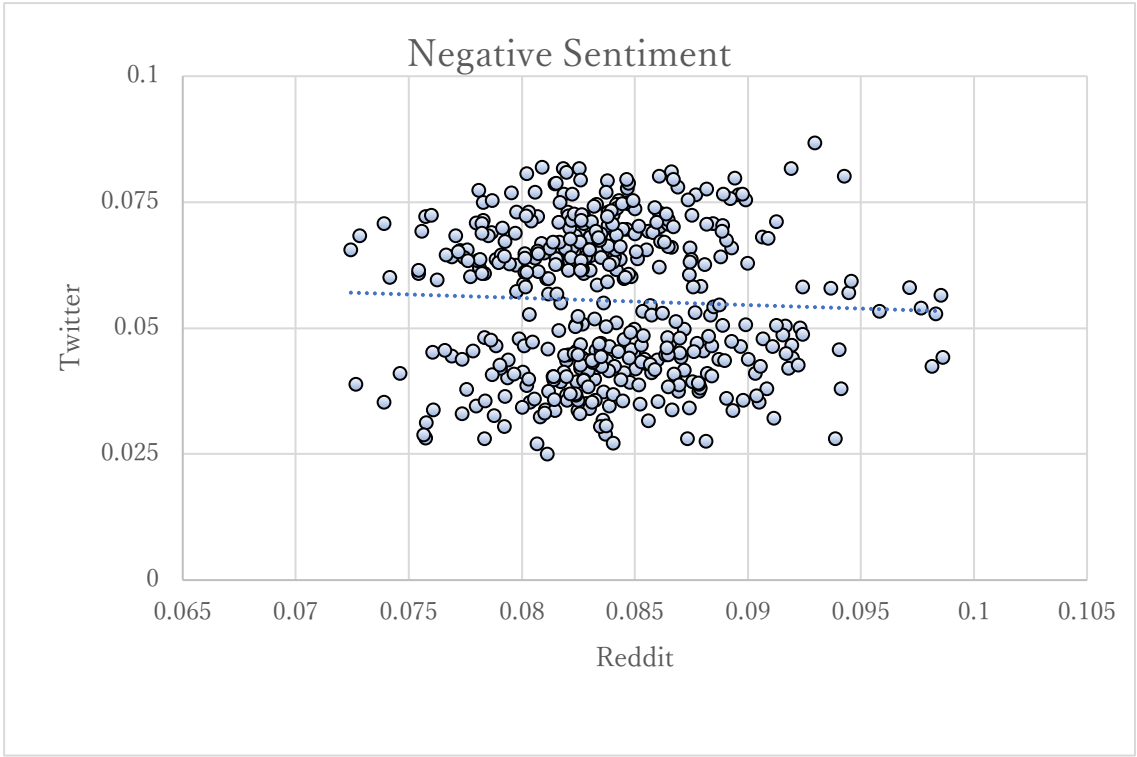


Figure 11: Twitter vs. Reddit Negative Sentiment Correlation (r = -0.042)

Figures 12-15 show the result of splitting the two datasets before and after May 2019; there is a higher correlation for both positive and negative sentiment after May 2019:

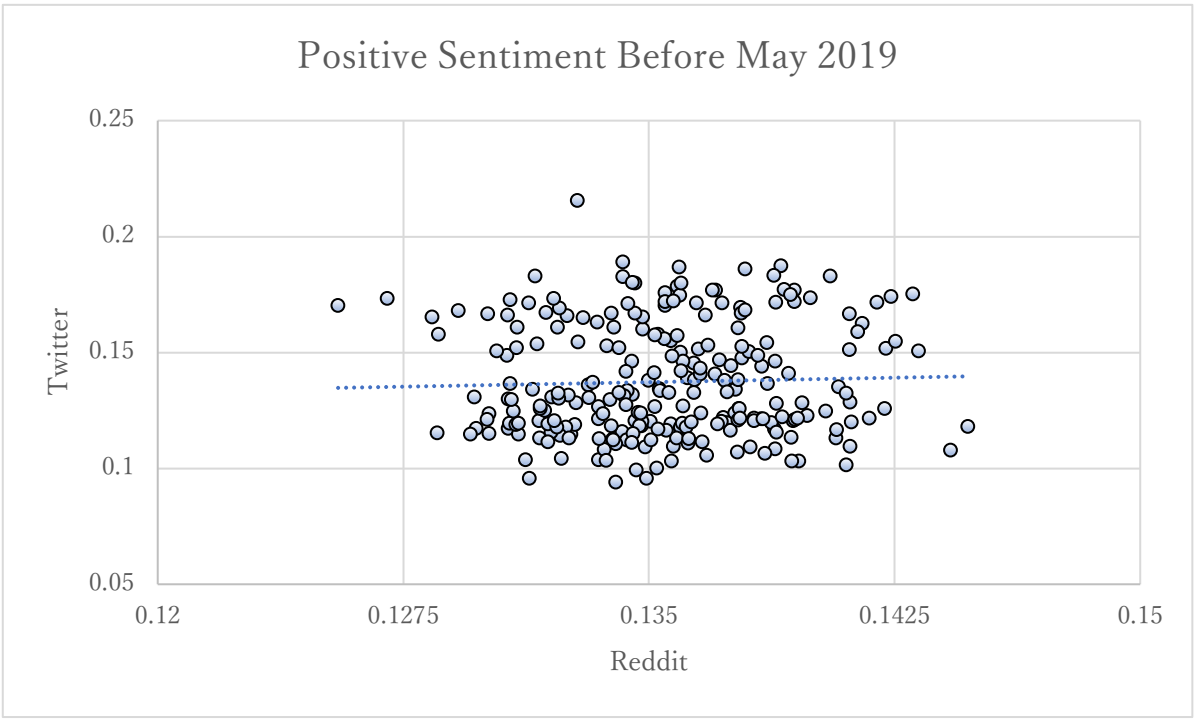


Figure 12: Twitter vs. Reddit Positive Sentiment Correlation; Data Before May 2019 (r = 0.037)

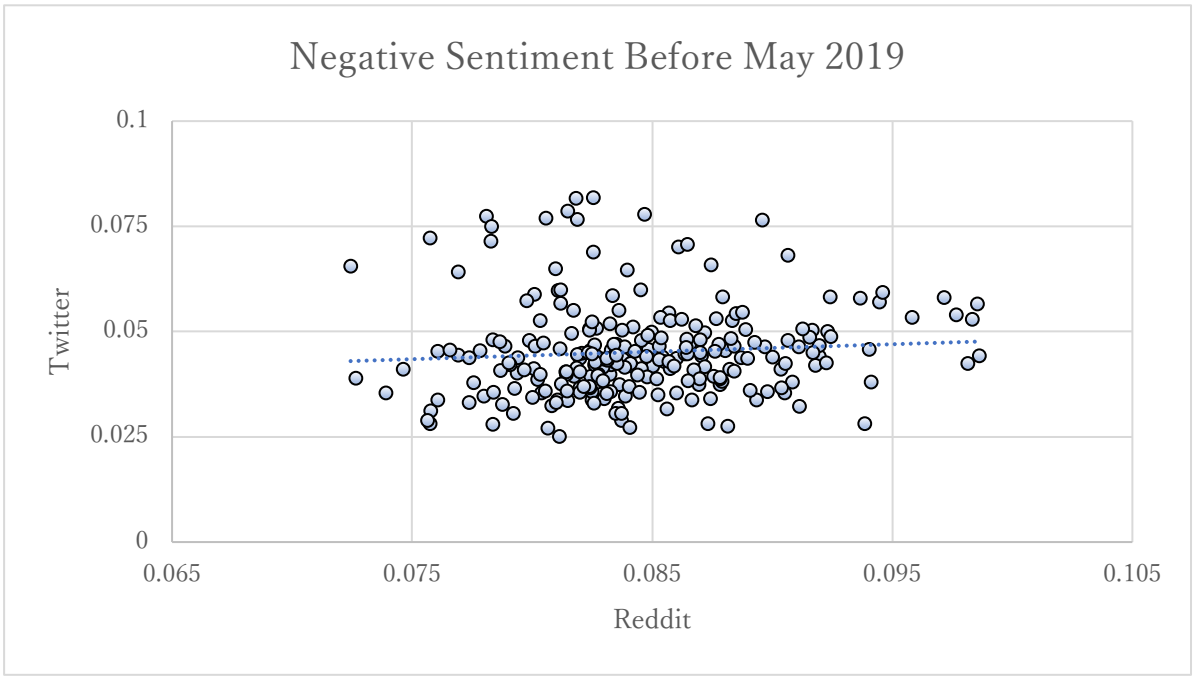


Figure 13: Twitter vs. Reddit Negative Sentiment Correlation; Data Before May 2019 (r = 0.079)

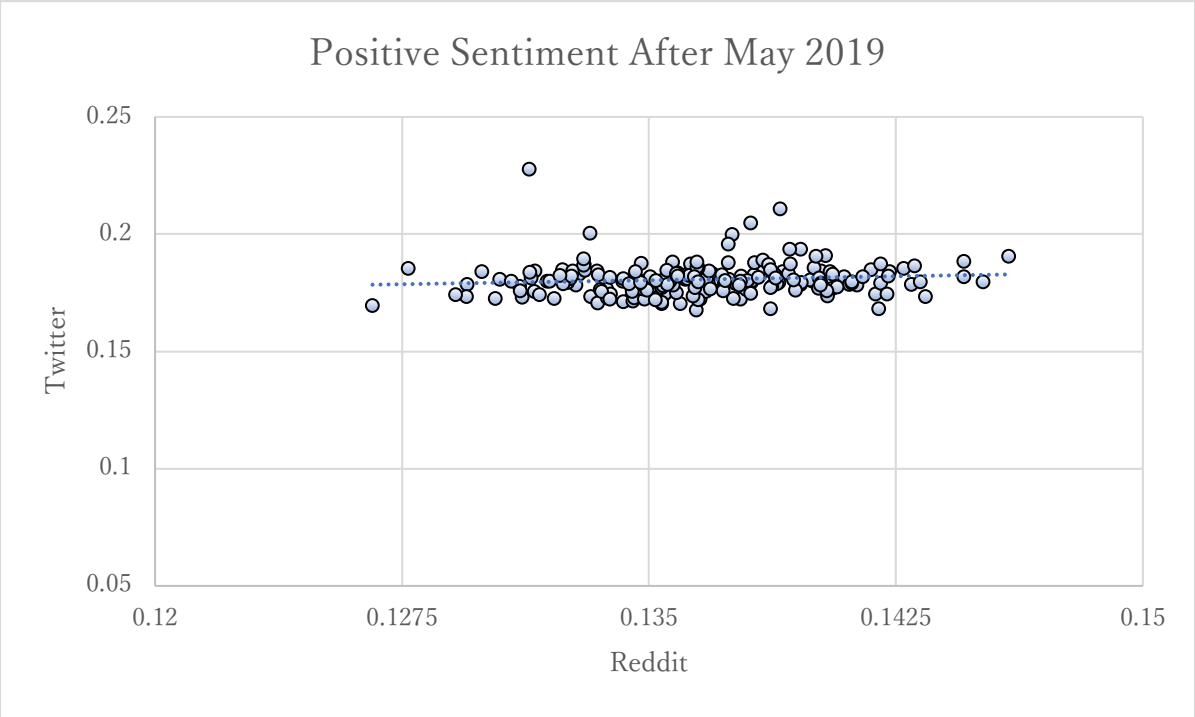


Figure 14: Twitter vs. Reddit Positive Sentiment Correlation; Data After May 2019 (r = 0.114)

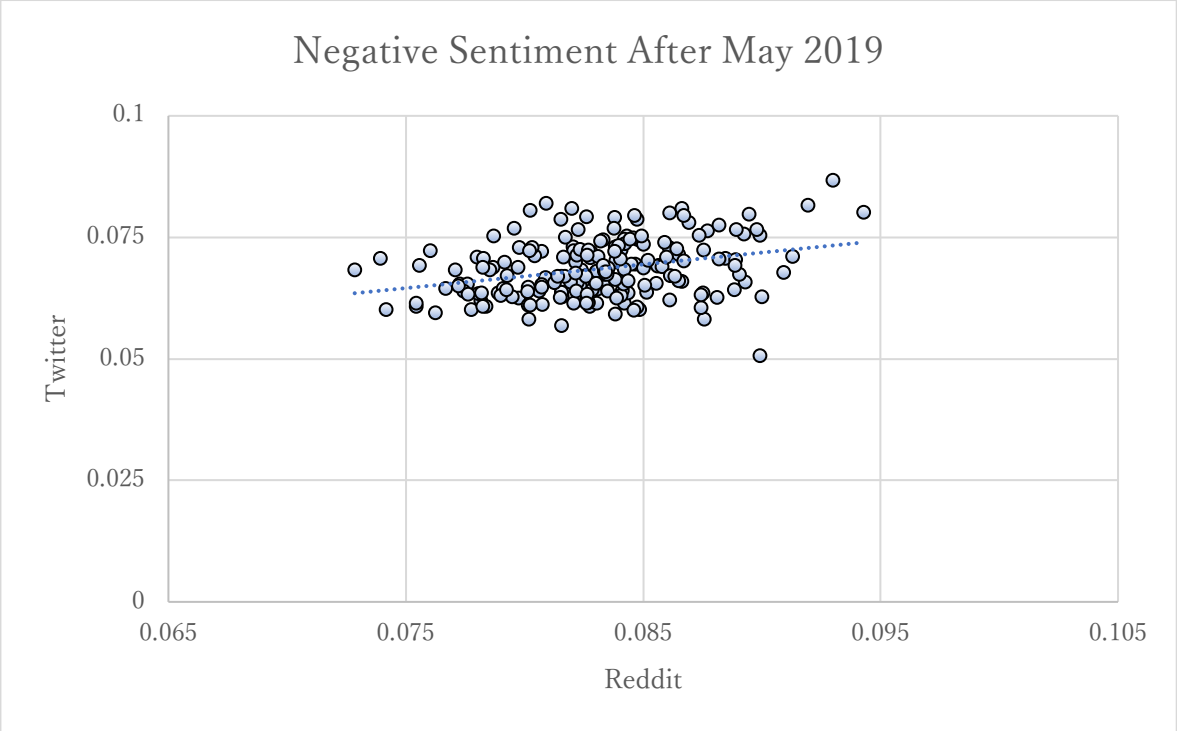


Figure 15: Twitter vs. Reddit Negative Sentiment Correlation; Data After May 2019 (r = 0.305)

Having two clusters can impact model accuracy and normalized features when input into the data. A way to handle this would be to split the datasets. However, following Critien et al. (2022), we do not split the data.

Machine Learning Models

CNN (Convolutional Neural Network), LSTM (Long-Short Term Memory), and BiLSTM (Bi-directional Long-Short Term Memory) deep machine learning models are used to predict both price direction and magnitude (multiclass models).

Convolutional Neural Networks (CNNs) are a type of neural network designed specifically for image and video recognition tasks. They are particularly effective because they can automatically learn features from the input data without human intervention. This makes them popular for object classification, object detection, and face recognition tasks. However, in recent years, CNNs have gained popularity in other fields, such as time-series problems like natural language processing and price predictions.

Long Short-Term Memory (LSTM) networks are Recurrent Neural Networks (RNN) that can capture long-term dependencies in sequence data. They are beneficial for tasks such as language translation, language generation, and speech recognition, where the context of a word or phrase is essential in understanding its meaning. In the case of price predictions, they can be helpful for remembering price movement trends in the past.

Bidirectional LSTM (BiLSTM) networks are a variant of LSTM networks that process the input sequence in both forward and backward directions, allowing the network to use both past and future context when making predictions. They are often used for tasks such as natural language processing, where the meaning of a word or phrase can depend on the context both

before and after it in the sentence. Just as in LSTM networks, they can be used in price predictions to remember previous price trends while also considering future trends.

We use the same models and notebooks as Critien et al. (2022), with only slight adjustments to the models to allow for the additional Reddit explanatory variables. Following Critien et al. (2022), instead of predicting a specific price, the magnitude is measured in percent increase/decrease bins (example: 0-10%, 10-20%, etc.), and the accuracy is measured using a confusion matrix (classification problem instead of regression).

The target variable bins seen in Figure 16 are used in the training dataset, the count being the number of times that target variable occurs in the training dataset.

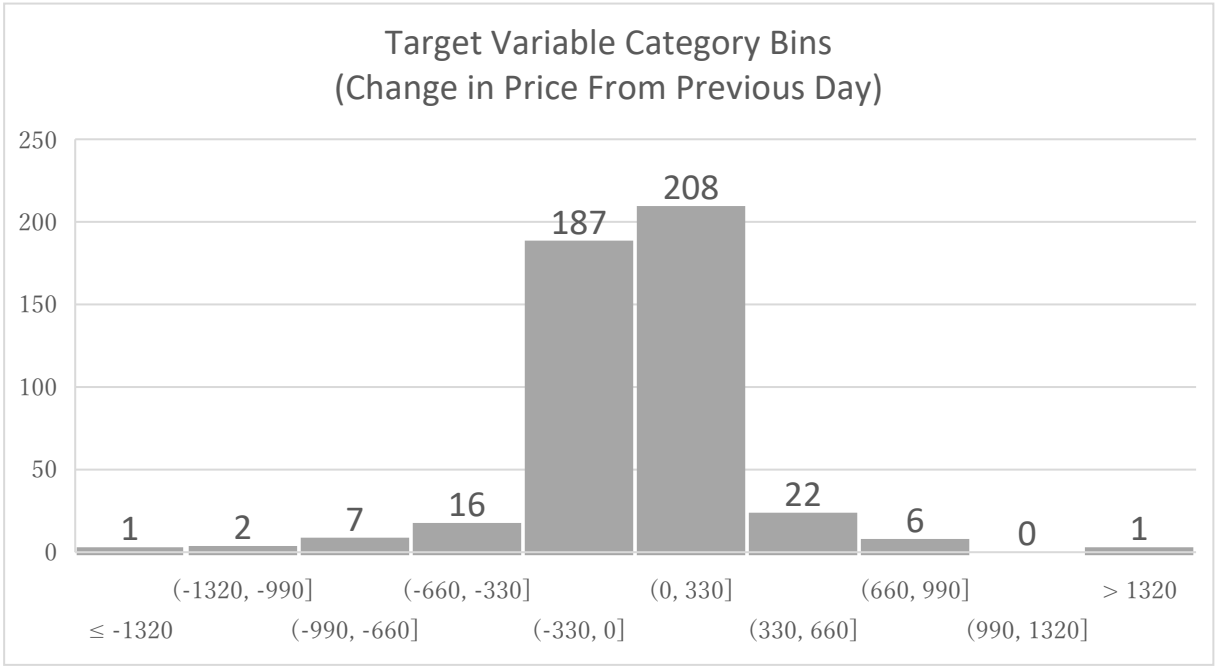


Figure 16: Number of Data Points in Each Target Variable Category Bin

Ideally, a balanced number of samples from each bin would be present in the dataset for better-generalized prediction accuracy. Critien et al. (2022) uses the dataset as is, so we follow in like manner.

Each dataset (1-, 3-, 7-day lag) is run through each model using a hyperparameter grid search, adjusting the following parameters:

1. Lagged Features

Selectable Hyperparameters: {1, 3, 7, 14} (4 Total)

Lagged features are derived from previous values of a time series. They are often used in time series modeling, particularly in forecasting future values of the time series. For example, if you are trying to predict future product demand, you might use lagged features, such as the demand for the product in the past few weeks or months, as input to your model. Lagged features can be helpful because they capture patterns in the time series that may be useful for predicting future values.

2. Neurons

Selectable Hyperparameters: {16, 32, 64, 128, 256} (5 Total)

In a neural network, a neuron receives input, processes it with a mathematical operation, and produces an output that is passed to other neurons in the network. These interconnected neurons can perform complex tasks such as image or speech recognition. The connections between neurons are weighted and adjusted during training, in which the network is presented with input-output pairs. The weights are updated through backpropagation to minimize error and allow the network to learn a specific task.

3. Layers

Selectable Hyperparameters: {1, 2, 3} (3 Total)

In a neural network, a layer is a group of neurons connected to the neurons in the preceding layer. The input layer receives the raw input data, and the output layer produces the network's final output. The layers in between are called hidden layers, and they extract features and patterns from the input data to transform it for use in making predictions. The number and

type of layers and the number of neurons in each layer play a significant role in the network's ability to learn and perform a specific task.

4. Batch Size

Selectable Hyperparameters: {5, 20, 50, 80} (4 Total)

Batch size is the number of samples processed together during training. The batch size can affect the speed and stability of the training process, with larger batch sizes leading to faster training but potentially less stability and smaller batch sizes leading to slower training but potentially more stability.

Grid search is a hyperparameter optimization method that involves training and evaluating a model for each possible combination of hyperparameter values specified in a grid. It is a brute-force approach that is simple to implement but computationally expensive. It is often used to find good hyperparameter values for a machine learning model.

The above grid search produced 240 (4 x 5 x 3 x 4 = 240) combinations of parameters for each dataset to be run through for each model. For each combination of parameters, the data runs through the model five times, randomly splitting the data with a 0.85-0.15 train-test ratio, with the max, min, and mean accuracies of the five runs recorded. The resulting 240 samples are used to compare results between the Twitter and Twitter-Reddit datasets.

5. Results

From a general review of the effect of adding Reddit data to Twitter data as an explanatory variable, Figures 17-19 show the differences in average mean and max accuracies of the 240, 5 run samples for each model-dataset pair.

Distribution of Mean and Max Accuracies by Model (BiLSTM)

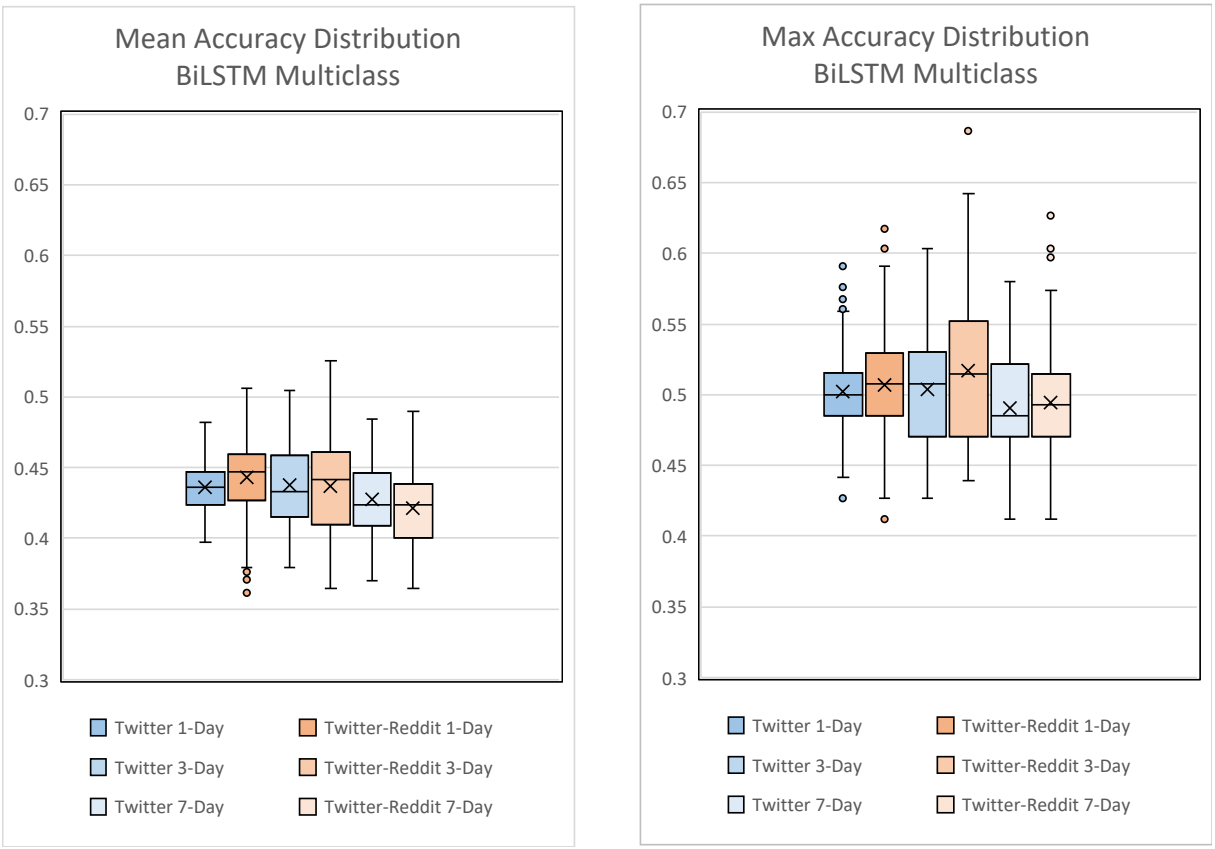


Figure 17: Distribution of Mean and Max Accuracies (BiLSTM)

The Twitter-Reddit datasets achieve a higher median, mean, and max accuracy in the mean and max accuracy distributions when compared against same-day lag datasets. However, there is a broader distribution compared to the Twitter datasets, suggesting a higher variance, and the model's results are not as stable when adding Reddit data.

Distribution of Mean and Max Accuracies by Model (CNN)

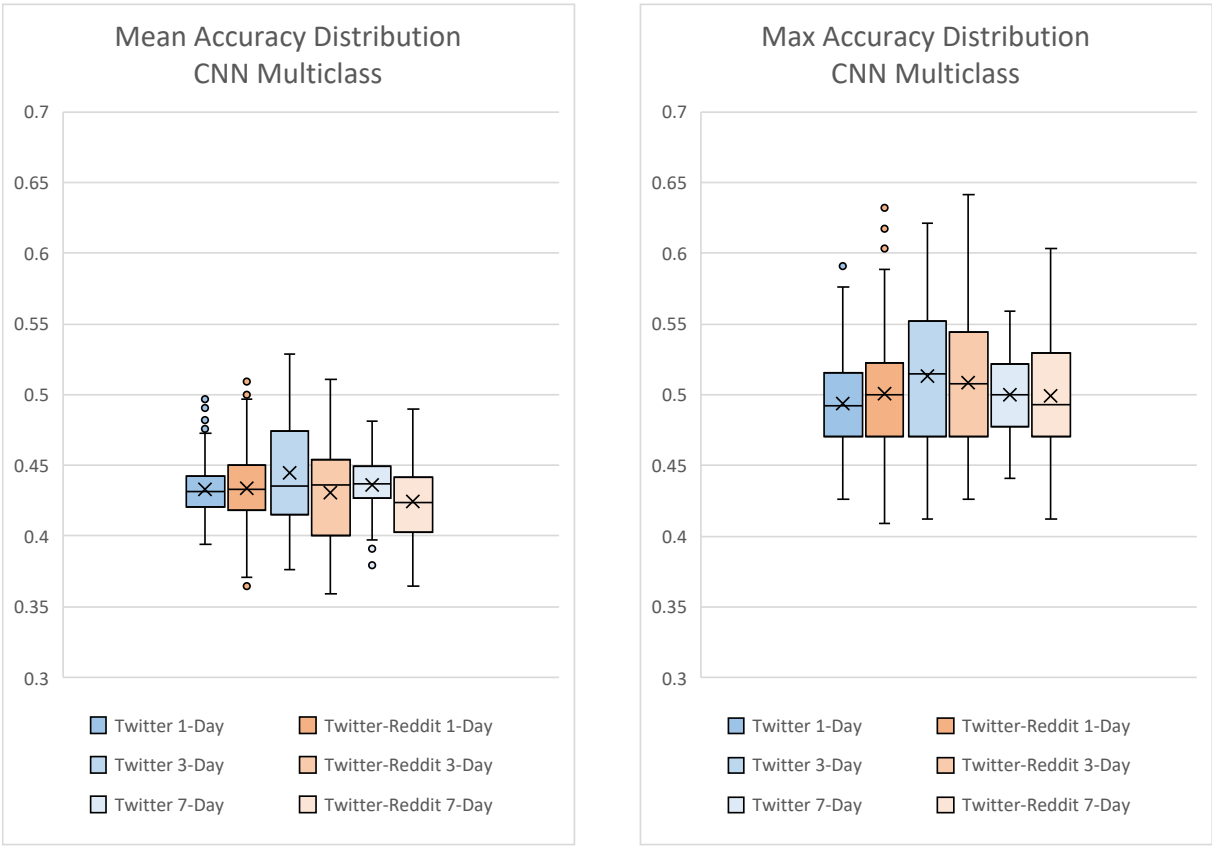


Figure 18: Distribution of Mean and Max Accuracies (CNN)

Adding Reddit data only sees a statistically significant increase in max accuracy when added to the 3-day lag dataset. Otherwise, there is either no change or a negative impact on accuracy when Reddit data is added to the dataset.

Distribution of Mean and Max Accuracies by Model (LSTM)

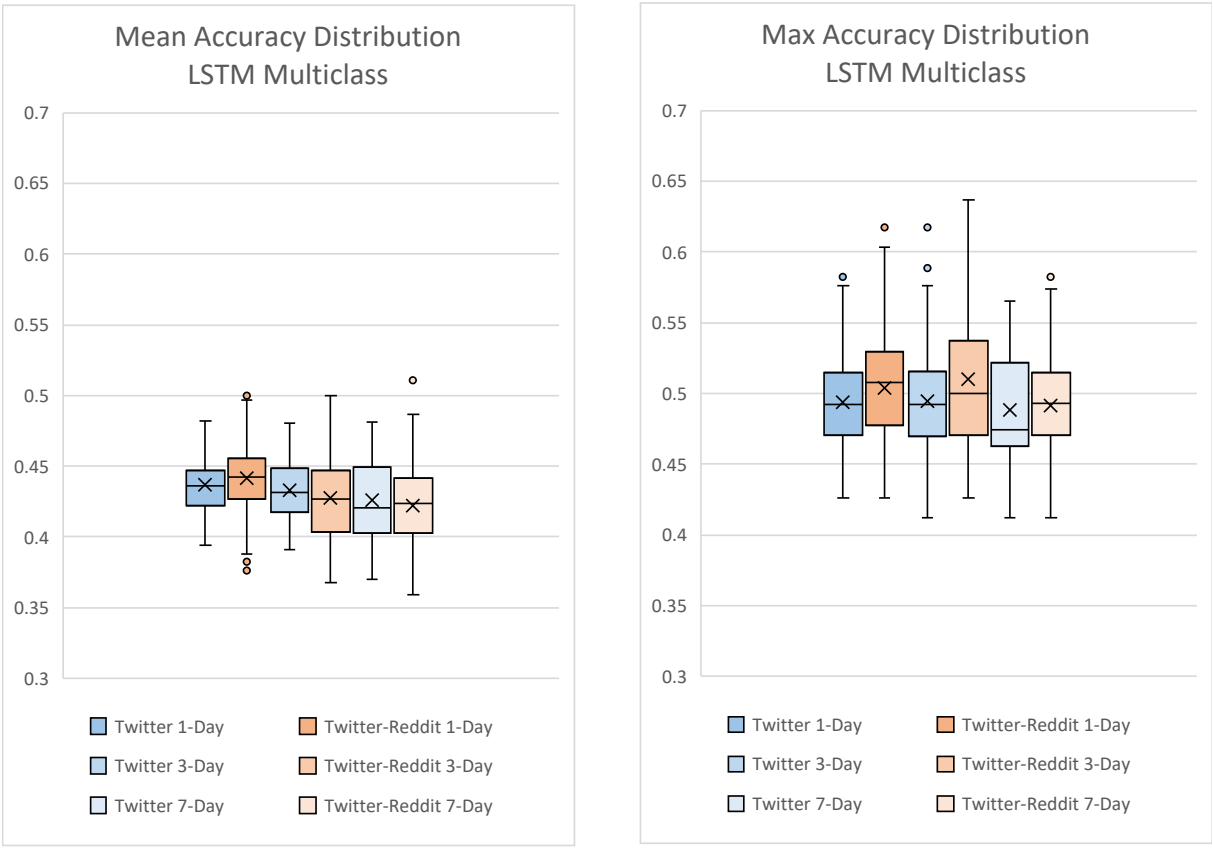


Figure 19: Distribution of Mean and Max Accuracies (LSTM)

Adding Reddit data to the LSTM models significantly increases the mean, median, and max accuracies for the 1-day dataset for mean accuracy and the 1- and 3-day datasets for max accuracy, but there is broader distribution. There is also a negative impact on 3- and 7-day datasets when considering mean accuracy.

Re-running the Simulation using Optimum Parameters from the Grid Search

Figure 20 shows the datasets and parameter combinations used to create the highest mean and max accuracies in the original sample datasets. These were then used in an attempt to achieve the highest accuracy and create a larger sample size to compare spread and test statistical significance. Ideally, accuracy results from each dataset, parameter, and model combination would be sampled more than 100 times to determine the optimal combination. Due to time constraints, we followed Critien et al. (2022) by selecting only the highest performing combinations from five samples.

Measurement	Highest Mean Accuracy		Highest Max Accuracy	
Dataset	Twitter	Twitter-Reddit	Twitter	Twitter-Reddit
Samples	5	5	5	5
Accuracy	0.528	0.525	0.621	0.687
Model	CNN	BiLSTM	CNN	BiLSTM
Day Lag	3	3	3	3
Lagged Features	7	7	14	7
Batch Size	5	5	5	20
Neurons	32	128	32	64
Layers	2	3	1	2

Figure 20: Highest Performing Parameters from Initial Simulation

Comparison of Simulation Results using Highest Mean Accuracy Parameters

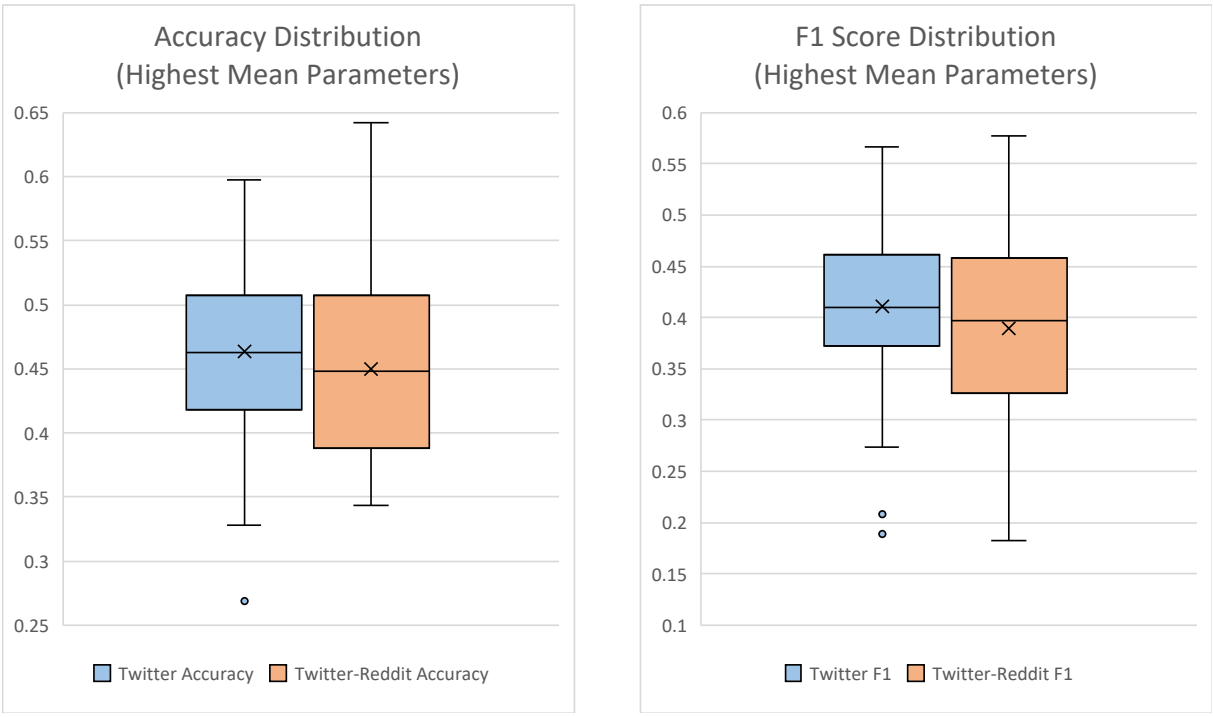


Figure 21: Accuracy and F1 Score Distributions (Highest Mean Accuracy Parameters)

t-Test: Accuracy Comparison (Highest Mean Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.463880597	0.450149254
Variance	0.003976859	0.004557583
Observations	100	100
Pooled Variance	0.004267221	
Hypothesized Mean Difference	0	
df	198	
t Stat	1.486364943	
P(T<=t) one-tail	0.069386598	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.138773196	
t Critical two-tail	1.972017478	

t-Test: F1 Score Comparison (Highest Mean Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.411195873	0.389741875
Variance	0.005634317	0.008160985
Observations	100	100
Pooled Variance	0.006897651	
Hypothesized Mean Difference	0	
df	198	
t Stat	1.826596534	
P(T<=t) one-tail	0.034633134	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.069266269	
t Critical two-tail	1.972017478	

Figure 22: Accuracy and F1 Score t-Tests (Highest Mean Parameters)

Although the Twitter-Reddit results achieve a higher max accuracy, the t-Test shows no statistical difference between the two model results. The F1 t-Test scores also suggest that the Twitter dataset used alone fits the data significantly more than when the Reddit data is added.

Comparison of Simulation Results using Highest Max Accuracy Parameters

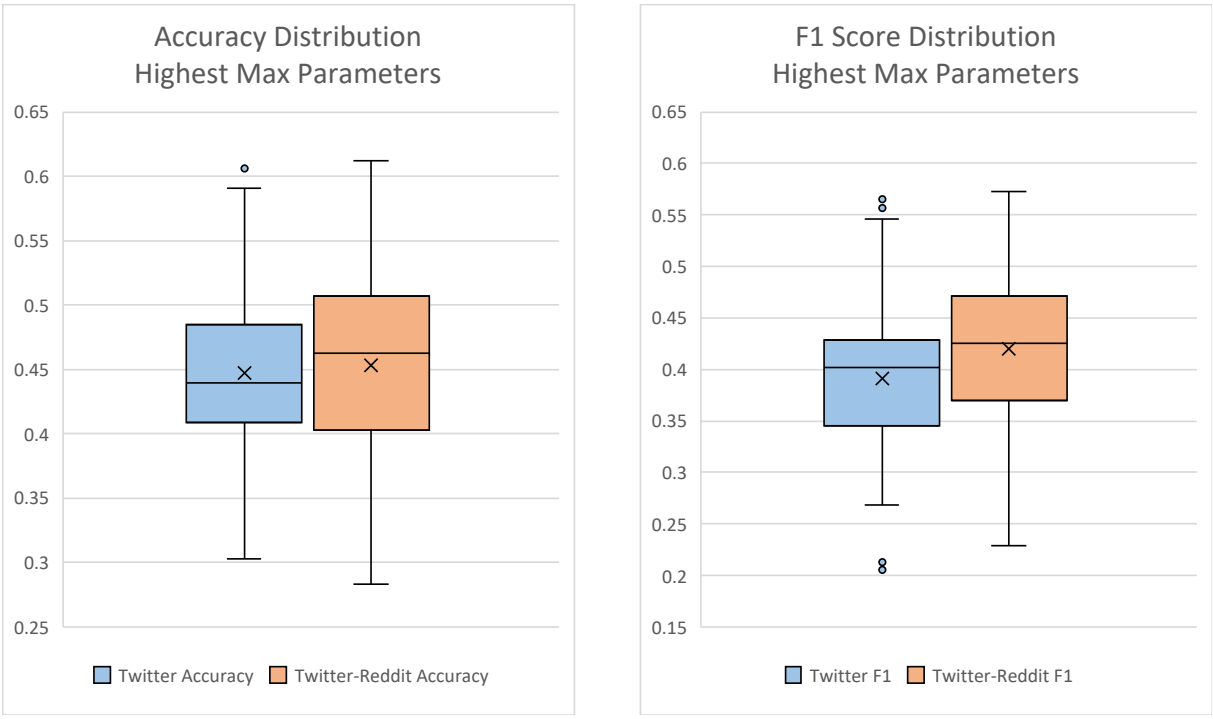


Figure 23: Accuracy and F1 Score Distributions (Highest Max Accuracy Parameters)

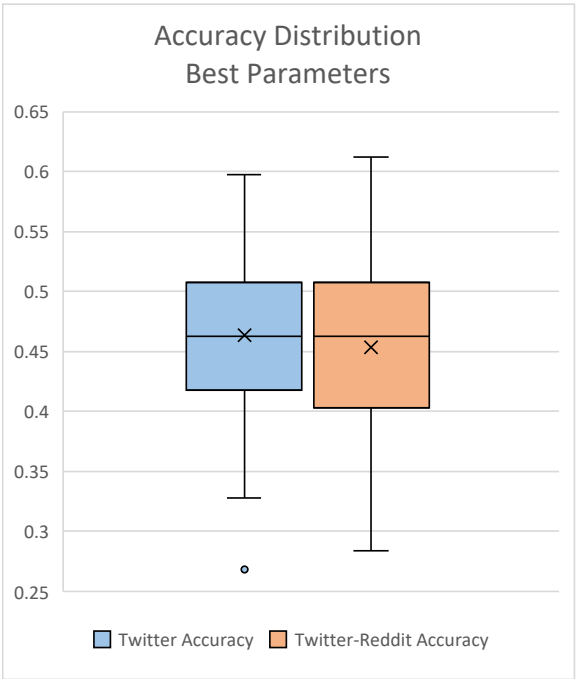
t-Test: Accuracy Comparison (Highest Max Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.447272727	0.453134328
Variance	0.003642856	0.005038219
Observations	100	100
Pooled Variance	0.004340538	
Hypothesized Mean Difference	0	
df	198	
t Stat	-0.629114202	
P(T<=t) one-tail	0.264999725	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.52999945	
t Critical two-tail	1.972017478	

t-Test: F1 Score Comparison (Highest Max Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.391353443	0.419742324
Variance	0.005239747	0.006254407
Observations	100	100
Pooled Variance	0.005747077	
Hypothesized Mean Difference	0	
df	198	
t Stat	-2.647950009	
P(T<=t) one-tail	0.004375338	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.008750675	
t Critical two-tail	1.972017478	

Figure 24: Accuracy and F1 Score t-Tests (Highest Max Parameters)

Although there is no statistical difference between the two accuracies, the t-Test of the F1 scores shows a significant increase in the F1 score for the Twitter-Reddit data. As there are conflicting results between the two, the best results from the Twitter and Twitter-Reddit datasets were compared below.

Comparison of Best Twitter and Twitter-Reddit Dataset Simulation Results



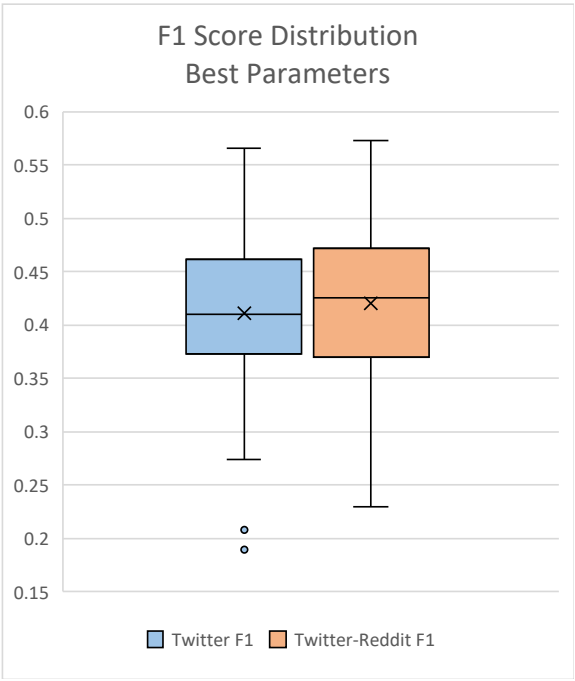


Figure 25: Accuracy and F1 Score Distributions (Highest Performing Parameters)

t-Test: Accuracy Comparison (Best Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.463880597	0.453134328
Variance	0.003976859	0.005038219
Observations	100	100
Pooled Variance	0.004507539	
Hypothesized Mean Difference	0	
df	198	
t Stat	1.131808471	
P(T<=t) one-tail	0.129541918	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.259083837	
t Critical two-tail	1.972017478	

t-Test: F1 Score Comparison (Highest Max Parameters)		
	Twitter	*Twitter-Reddit*
Mean	0.411195873	0.419742324
Variance	0.005634317	0.006254407
Observations	100	100
Pooled Variance	0.005944362	
Hypothesized Mean Difference	0	
df	198	
t Stat	-0.783823356	
P(T<=t) one-tail	0.217040204	
t Critical one-tail	1.652585784	
P(T<=t) two-tail	0.434080407	
t Critical two-tail	1.972017478	

Figure 26: Accuracy and F1 Score t-Tests (Highest Max Parameters)

Once again, there was no statistical difference between the two results, suggesting that adding Reddit data to any model (BiLSTM, CNN, and LSTM) would not increase the highest achievable mean accuracy.

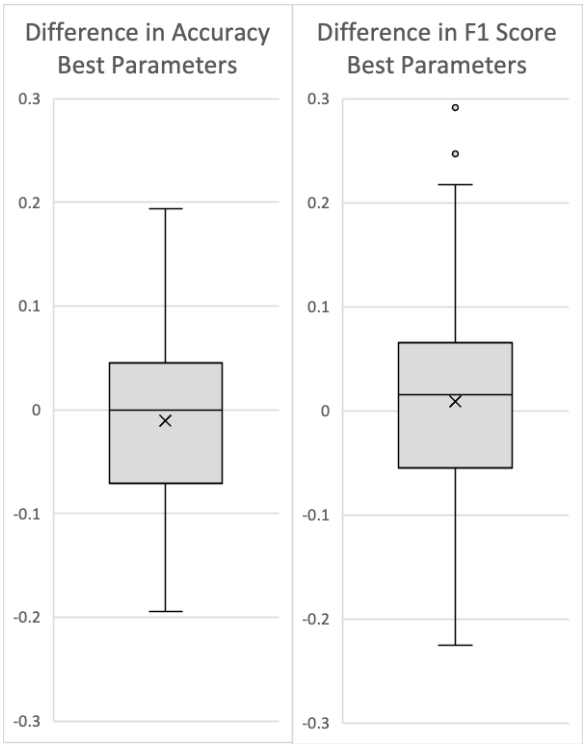


Figure 17: Distribution of the Difference in Accuracy and F1 Score per Run

We also see from Figure 27 that the average and mean differences in accuracy and F1 score per run are close to zero, suggesting that even though there are some runs achieving a much higher accuracy or F1 score, it is balanced out by the decreases in accuracy and F1 score.

6. Conclusion

This research used machine learning models to predict the direction and magnitude of bitcoin price movements using data from Twitter and Reddit. When considering the distribution of mean and max accuracy results, adding Reddit data increased the spread of the distribution, indicating higher variance in the results. The study also found that, in some cases, adding Reddit data resulted in a statistically significant increase in the F1 score. Still, there was no statistical difference in mean or max accuracy between the models using Twitter data alone and those using both Twitter and Reddit data.

As a note, we aimed to predict the actual price change of bitcoin rather than its volatility. Price volatility has been researched extensively within the academic community. Still, it has been excluded from the scope of our specific research, as we hoped to make progress in predicting price change. Due to its importance in the financial econometrics literature, further research in volatility forecasting should also be pursued in the future.

We primarily focused on improving prediction accuracy by combining Reddit data with Twitter data, but the inclusion of news text data as an explanatory variable could further improve accuracy. News text data, such as news articles and press releases, have been shown to contain valuable information that can be used to predict financial outcomes. News often goes through different channels and may be disseminated to investors faster compared to posts on Reddit or Twitter. There are times when social media posts become the base content of news, like Elon Musk's posts, but more often than not news is the base content for social media posts. As such, news text data could also be a leading indicator of Bitcoin price movement and magnitude over social media data, and incorporating news text data as an explanatory variable to increase prediction accuracy is a valuable research topic and could provide more actionable insights for practitioners.

This research also noted potential improvements in the application of machine learning techniques within the research framework conducted by Critien et al. (2022). Some of those were clusters of data not being separated for training and validation (before and after May 2019), categories not being equally distributed in the training data, and selecting parameters that created the highest max accuracy and highest mean accuracy without consideration for sample size (only five samples were parameter combination) or variance. Future research would need to consider these potential improvements.

Other research opportunities would include grouping data and predicting on an hourly interval instead of daily, extending research to include the 2021 cryptocurrency bubble period, and taking all Reddit comments from subreddits that are highly correlated with cryptocurrency price movement (Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S., 2019) while this research only considered Reddit comments that included "Bitcoin", as well as researching the optimal lag combination of both Twitter and Reddit data as this research put Twitter and Reddit on the same daily lag intervals. In essence, this study treated and processed Reddit data the same way as Twitter data, which ended in no significant increase in accuracy. As the characteristics of the two platforms differ, it would be best for future research to consider these differences and process Reddit data more accordingly to be considered a valuable explanatory variable.

As a practical implication of this research, especially from the interests of investors, it is important to note that although the accuracy scores of these predictive neural networks can reach up to and beyond 70% max accuracies, the variance in accuracy is too high for these models to be considered a stable option. It is suggested that investors hoping to utilize these models for investment purposes take careful note of how accuracy scores are being reported (whether mean, median, or max accuracy), how that accuracy is calculated (how many

samples were used), and what data the model is trained on (date intervals, variables, etc.) before using them in an official setting.

Acknowledgements

I would like to express my sincere gratitude to Professor Takaki Hayashi for his technical advice and thoughtful direction. His extensive knowledge and expertise in the field of statistics and machine learning have been invaluable to the development of this research, as well as my personal development as a researcher. I am highly grateful for his expertise, patience, time, and support.

I would also like to extend my heartfelt thanks to Professors Sachiko Yamao and Fumiko Takeda for their insights and suggestions. Their sound advice was critical in refining the direction of this research and the overall quality of the research.

References

Abraham, J., Higdon, D., Nelson, J. & Ibarra, J. (2018) 'Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis', *SMU Data Science Review*: Vol. 1: No. 3, Article 1 [online]. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/1/ (Accessed: 22 June 2022)

Alex. (2019, November 23). *Bitcoin tweets - 16m tweets*. Kaggle. Available at: https://www.kaggle.com/datasets/alaix14/bitcoin-tweets-20160101-to-20190329 (Accessed: 2 January 2023)

Bremmer, G. (2018) 'Predicting tomorrow's cryptocurrency price using a LSTM model, historical prices and Reddit comments', *Tilburg University* [online]. Available at: http://arno.uvt.nl/show.cgi?fid=147279 (Accessed: 22 June 2022)

C Phillips, R., & Gorse, D. (2018) 'Predicting cryptocurrency price bubbles using social media data and epidemic modelling', *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1-7, doi: 10.1109/SSCI.2017.8280809 [online]. Available at: https://ieeexplore.ieee.org/document/8280809 (Accessed: 22 June 2022)

Fanelli, J. (2020, July 18). *Reddit comments containing "bitcoin" 2009 to 2019*. Kaggle. Available at: https://www.kaggle.com/datasets/jerryfanelli/reddit-comments-containing-bitcoin-2009-to-2019 (Access: 2 January 2023)

'Forecasting Cryptocurrency Price Movements with Deep Learning and Sentiment Analysis', *University of Warwick* [online]. Available at: https://warwick.ac.uk/fac/soc/economics/current/modules/ec331/raeprojects/ec331_-_1610710_-_ec331_dissertation.pdf (Accessed: 22 June 2022)

Gurrib, I. & Kamolov, F. (2021) 'Predicting bitcoin price movements using sentiment analysis: a machine learning approach', *Studies in Economics and Finance* [online]. Available at: https://www.emerald.com/insight/content/doi/10.1108/SEF-07-2021-0293/full/html (Accessed: 22 June 2022)

H Alejandro, R. (2021), 'Twitter and Reddit posts analysis on the subject of Cryptocurrencies' [online]. Available at: https://www.researchgate.net/publication/352262235_Twitter_and_Reddit_posts_analysis_on_the_subject_of_Cryptocurrencies (Accessed: 22 June 2022)

Hutto, C.J. 'vaderSentiment' [online]. Available at: https://github.com/cjhutto/vaderSentiment (Accessed: 22 June 2022)

Kraaijeveld O, De Smedt J (2020) The predictive power of public twitter sentiment for forecasting cryptocurrency prices. J Int Finan Markets Inst Money 65:101188. Available at: https://doi.org/10.1016/j.intfin.2020.101188 (Accessed: 2 January 2023)

Lamon, C., Nielson, E. & Redondo, E. (2017) 'Cryptocurrency Price Prediction Using News and Social Media Sentiment' [online]. Available at: https://www.semanticscholar.org/paper/Cryptocurrency-Price-Prediction-Using-News-and-Lamon-Nielsen/c3b80de058596cee95beb20a2d087dbcf8be01ea (Accessed: 22 June 2022)

Critien, J.V., Gatt, A. & Ellul, J. (2022) 'Bitcoin price change and trend prediction through twitter sentiment and data volume', *Financ Innov*, 8, 45 [online]. Available at: https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00352-7 (Accessed: 22 June 2022)

Marne, S., Churi, S., Correia, D. & Gomes, J. (2021), 'Predicting Price of Cryptocurrency – A Deep Learning Approach', *International Journal of Engineering Research & Technology (IJERT)*, ISSN: 2278-0181 [online]. Available at: https://www.ijert.org/research/predicting-price-of-cryptocurrency-a-deep-learning-approach-IJERTCONV9IS03083.pdf (Accessed: 22 June 2022)

Mohapatra S, Ahmed N, Alencar P (2020). KryptoOracle: a real-time cryptocurrency price prediction platform using twitter sentiments. Available at: arXiv:2003.04967 (Accessed: 2 January 2023)

Reid, D (2019, May 14). *Bitcoin passes $8,000 as value more than doubles in 2019*. CNBC. Available at: https://www.cnbc.com/2019/05/14/bitcoin-passes-8000-as-value-more-than-doubles-in-2019.html (Accessed: 5 January 2023)

Sawhney, R., Agarwal, S., Mittal, V., Rosso, P., Nanda, V., & Chava, S. (2022, May 11). *Cryptocurrency bubble detection: A new stock market dataset, Financial Task & Hyperbolic models*. arXiv.org. Available at: https://arxiv.org/abs/2206.06320 (Accessed: 2 January 2023)

Serafini G, Yi P, Zhang Q, Brambilla M, Wang J, Hu Y, Li B (2010) Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches. In: 2020 International joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19–24, 2020, pp. 1–8. IEEE (2020). Available at: https://doi.org/10.1109/IJCNN48605.2020.9206704 (Accessed: 2 January 2023)

Stevekovach. (2021, February 8). *Tesla buys $1.5 billion in bitcoin, plans to accept it as payment*. CNBC. Available at: https://www.cnbc.com/2021/02/08/tesla-buys-1point5-billion-in-bitcoin.html (Accessed: 30 December 2022)

Tandon, C., Revankar, S., Palivela, H. & S Parihar, 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*, Volume 1 (Issue 2) [online]. Available at:

https://www.sciencedirect.com/science/article/pii/S2667096821000288# (Accessed: 22 June 2022)

Throuvalas, A. (2022, December 20). *Portion of bitcoin supply held by retail investors reaches all-time high: Glassnode*. Decrypt. Available at: https://decrypt.co/117685/portion-bitcoin-supply-held-retail-reaches-all-time-high-glassnode (Accessed: 30 December 2022)

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019, June 14). *Price movement prediction of cryptocurrencies using sentiment analysis and machine learning*. MDPI. Available at: https://doi.org/10.3390/e21060589 (Accessed: 2 January 2022)

Wooley, S., Edmonds, A., Bagavathi, A. & Krishnan, S. (2019) 'Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment', *IEEE* [online]. Available at: https://ieeexplore.ieee.org/document/8999092 (Accessed: 22 June 2022)

Zaman, S., Yaqub, U. and Saleem, T. (2022), "Analysis of Bitcoin's price spike in context of Elon Musk's Twitter activity", *Global Knowledge, Memory and Communication*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/GKMC-09-2021-0154 (Accessed: 30 December 2022)

Zielak. (2021, April 11). *Bitcoin historical data*. Kaggle. Available at: https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data (Accessed: 2 January 2023)