Universiteit Leiden
The Netherlands

Opleiding Informatica

Using sentiment analysis on tweets for predicting cryptocurrency price movements: A case study

Rutger H.J Klaassen

Supervisors:
Jan N. van Rijn & Yash J. Patel

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/08/2023

Abstract

Cryptocurrencies have gained significant relevance in the financial landscape and are increasingly viewed as an emerging market. The cryptocurrency market's characteristics, mainly the abundant availability of different types of market data, make it a compelling and suitable subject for research. By employing sentiment analysis and machine learning techniques, one can explore valuable insights into market behaviour, thereby addressing the complex task of stock market prediction.
This research investigates the effectiveness of different methodologies for predicting cryptocurrency price movements, utilizing machine learning models in combination with data from sentiment analysis, technical analysis, or a hybrid approach merging both methods. In this case study, we focus on gathering data from social media platforms, specifically Twitter, and built classifiers that predict whether a cryptocurrency price will increase or decrease in price in the next time window (using hourly intervals) based on sentiment analysis, technical analysis, and a combination of both. Using these classifiers we investigated four different cryptocurrencies: Bitcoin, Ethereum, Dogecoin and Ripple.
The results reveal an average accuracy of 50.05% for sentiment analysis, 51.61% for technical analysis, and 51.46% for the combination of both. The majority class classifier obtains an average accuracy of 50.41%. A statistical significance test shows that sentiment analysis alone does not consistently outperform the baseline, while models incorporating technical analysis or the combination of both show promise in capturing market dynamics. We also investigated the effect of punctuation in text for the sentiment analysis and concluded based on our results that utilizing punctuation is a crucial factor for obtaining good results from sentiment analysis.

Contents

1 Introduction

In this section, we will provide an overview of cryptocurrencies, their current state, our contribution to the research field, our research question, and an outline of the thesis.

1.1 Cryptocurrencies

The first tradable cryptocurrency, Bitcoin, was introduced in 2009 by an anonymous person with the pseudonym Satoshi Nakamoto [1]. Since then thousands of cryptocurrencies have been created, each with its own features and use cases. As of 06-04-2023, Coinmarketcap [2] reports that there are approximately 23,217 different cryptocurrencies with a total market cap of $1.18 trillion. Among these, Bitcoin holds the largest market share, accounting for 45.8% of the total market capitalization, with Ethereum coming in second accounting for a 19.1% market capitalization. Cryptocurrency prices can be considered volatile if we look at the price. Ethereum for example, had an opening price of $0.74 in 2015 and peaked at $4,891.70 in 2021, see Figure 1.

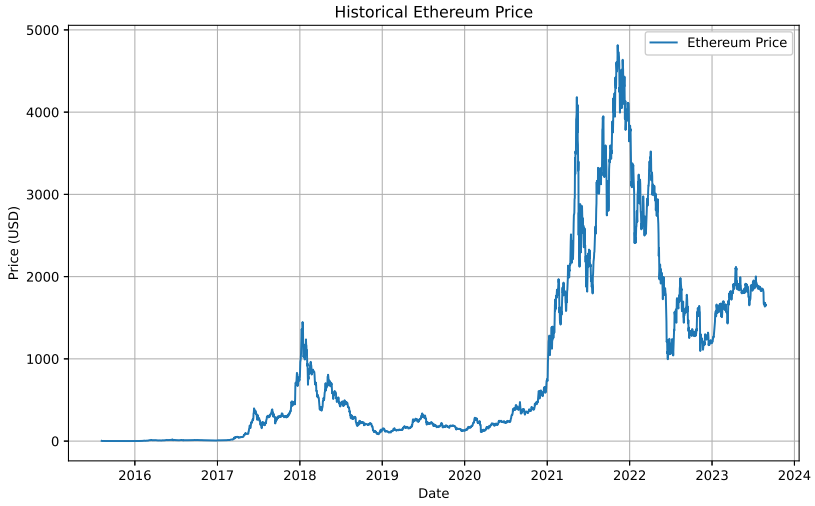

Figure 1: Ethereum price movements 2015 until 2023

1.2 The situation

Trading on financial markets presents both significant challenges and opportunities. To execute a successful trading strategy, various components must work harmoniously, such as identifying trends, evaluating risks involved, and predicting whether the price will go up or down. The latter is something that can be done by many different types of prediction models. A lot of these prediction models use technical analysis to forecast market trends [3]. Technical analysis is a method used to evaluate and forecast future price trends based on historical market data. Technical analysts use metrics derived from open, high, low, close & volume data, such as moving averages, relative strength index, moving average convergence divergence, and Bollinger bands, among others, to interpret the historical data and gain insights into the potential future price direction.
Cryptocurrency markets seem even more nascent than the regular stock market, which makes them interesting for technical analysis [4]. These markets are decentralized, allowing for transactions to occur 24 hours per day and 7 days per week, without the need for intermediaries such as banks or financial institutions. Without these intermediaries, cryptocurrency prices are determined solely by the supply and demand of buyers and sellers.

Social media is a platform where opinions about certain cryptocurrencies are voiced, and it has been demonstrated previously that there is a correlation between social media trends and cryptocurrency

price trends [5]. It is important to realize that there is a high degree of sentiment related to what people post. Whether they write positively or negatively about a cryptocurrency could possibly predict the price [6]. Sentiment can be determined by natural language processing as shown in previous research [7], which is what we will be using in our research as well.

1.3 Research question

This thesis will mainly focus on using machine learning to determine whether social media is a better indicator of price trends than the more used technical indicators based on price history. The research question that we try to answer is: *Is there a correlation between media sentiment analysis and cryptocurrency price movements, and how does it compare to technical analysis?*.

1.4 Our contribution

This research functions as a case study, where it expands on previously found techniques [6, 8] of determining cryptocurrency price movements using Twitter (presently known as 'X') sentiment analysis by placing it in a real-life context. By utilizing data from four different coins we aim to gain insights into the correlation between the sentiment voiced on social media and the price movements of cryptocurrencies. Additionally, this thesis analyzes the comparative performance of our method, which is machine learning based solely on sentiment analysis, against machine learning approaches based on technical analysis by employing technical indicators as input for creating a classification model and also investigate whether adding sentiment analysis to technical analysis gives favourable results. Finally, we also investigate the impact of punctuation on sentiment analysis.

1.5 Thesis overview

In Section 2 we will explain the related work in the domain of sentiment and technical analysis; Section 3 discusses the type of data used in this thesis and how we acquired it; Section 4 explains our methodology and Section 5 the experimental setup. Section 6 shows the results of our experiments. Section 7 discusses our research and shows the limitations. We conclude in Section 8 and describe future research in Section 9.

2 Related work

There has been a significant amount of research conducted on predicting financial markets for profit, particularly with the rising popularity of cryptocurrencies. In this section, we will discuss previous research that has inspired this thesis and explain our contributions to this specific field of research. This section also introduces the terms we used throughout the paper.

2.1 Technical analysis

The cryptocurrency market is nascent, making it hard to predict. However, multiple different strategies have been discovered that can predict price movements using technical analysis.
Long short-term memory network has been successfully applied to predict (describe what they predict), and obtained an accuracy of 52.76% [9]. This research used technical indicators derived from price data like the simple moving average and de-noised closing price, but also indicators that come from information about bitcoin itself like the mining difficulty and the hash rate.
Research has been done towards bitcoin price forecast on various time scales, ranging from one-day windows to ninety-day windows [10]. Using long short-term memory networks they managed to achieve 62-65% accuracy for their classification models. They also created a regression model with an error of 1.7% for a one-day time frame and an error between 2.88% to 4.10% for a seven to ninety-day time frame.
Technical analysis based on moving averages can yield favourable results in the cryptocurrency markets [4]. The study specifically utilizes the variable-length moving average strategy in one-minute intervals. This means that this strategy involves buying or selling based on the moving average every minute. The research concludes that it is possible to generate profits in the cryptocurrency market using this strategy. Techniques based on moving averages have been utilized for decades [11] and remain applicable in newer markets, as evidenced by Corbet et al.
Research has been conducted by employing random forests and hyperparameter optimization based on technical analysis [12]. Van der Avoird uses this technique to achieve a classification accuracy of roughly 53% on an hourly interval. In turn, his results could be used to retrieve a return on investment of 17.1% over a period of 15 months.

2.2 Price prediction using sentiment analysis

It has also been shown that the general sentiment of particular internet sources can predict price movements to a certain degree [6, 8]. As previously discussed, we will employ sentiment analysis on (social) media to predict price fluctuations. Sentiment analysis is a relatively new method used in various ways [13, 14]. Building on the techniques described in [13], we will conduct sentence-level analysis on social media-related posts. It has been shown that sentiment analysis can be used to predict cryptocurrency prices [6]. They collected cryptocurrency-related tweets through the Twitter API and performed sentiment analysis on these tweets. Sentiment was measured using valence sentiment analysis, which quantifies the level of positivity or negativity associated with an emotional experience. The study utilized the VADER sentiment analysis library [15], an open-source tool validated by humans and well-suited for analyzing social media content. The sentiment data was then fed into a support vector machine, random forest, and multi-layer perceptron to determine whether to buy or sell. By combining Twitter data with market data, they achieved a maximum accuracy of 72% using multi-layer perceptron for Bitcoin, indicating that the model accurately predicted price movements 72% of the time. Notably, the random forest model yielded satisfactory results for technical analysis but not for sentiment analysis or the combination of the two, resulting in an accuracy of only 44%.

Ethereum prices have been accurately predicted before using sentiment analysis on social media and news articles [8]. They employed support vector machines and long short-term memory networks, combining price history and sentiment analysis. The long short-term memory model, which integrated price history with sentiment analysis, outperformed the support vector machine model with a mean absolute normalized error of 1.36%.

Feature vectors combined with the naive-Bayes algorithm have also been used on sentiment analysis [16]. This approach achieved a day-to-day classifier accuracy of 55%. It is worth mentioning that they also attempted to use a support vector machine, but it consistently achieved lower accuracy compared to the naive-Bayes algorithm.

3 Data

In this section, we provide information about how we obtained the data from Twitter and financial repositories, cleaned the data, and modified the data to fit the framework.

3.1 Definitions

3.1.1 Open, Close, High, Low

For each given period, fixed at one hour for this thesis, four distinct price movement indicators track changes in currency value over time. We utilize this data to identify the technical indicators and price movements that serve as the basis for evaluating the accuracy of our machine-learning model. The four indicators are:

- **Open**: The cryptocurrency's price at the start of the time interval.
- **Close**: The cryptocurrency's price at the end of the time interval.
- **High**: The highest price at which the cryptocurrency has traded during the time interval.
- **Low**: The Lowest price at which the cryptocurrency has traded during the time interval.

3.1.2 Volume

In trading, volume refers to the amount of units a particular financial asset has been traded within a given period. Volume is an important metric in trading as it reflects the level of market activity and liquidity. Volume can be used to identify trends and potential price movements, which is especially useful in technical analysis.

3.2 Data acquisition

3.2.1 Twitter data

For this research, we used Twitter data to create our own data set. For this, we used the (now depleted) Python package, Twint [17]. Twint is a powerful Twitter tool that allows you to gather tweets from Twitter without requiring access to Twitter's API. One of the most common use cases of Twint is to gather data on specific topics or keywords, including cryptocurrencies. To do this, you can use the "search" functionality of Twint to specify the keywords and the time range of interest.

For collecting tweets about cryptocurrencies between a specific date and time, you can use the following command in the terminal:

twint -s "cryptocurrency" –since "2022-01-01 00:00:00" –until "2022-01-31 23:59:59" -o tweets.csv –csv

This command will search for tweets containing the word "Bitcoin" between January 1st, 2022, and January 31st, 2022, and save the results in a CSV file named "tweets.csv". You can modify the search query and the time range as per your requirements. Twint also had a language filter so we would only retrieve English tweets. This is necessary, as the natural language processing library we use to perform sentiment analysis could only perform accurate sentiment analysis on English texts.

3.2.2 Other social media data sources: Reddit, Youtube, news stories & blogs

For gathering data from Reddit we used the pushshift [18] data set. It is notable that as of May 2023, the pushshift data set has become unusable due to a ban from Reddit. The problem with Reddit is the small amount of available data, specifically when filtering on time and keywords. The platform's official API doesn't let you search the entire website, but only specific communities called subreddits. This feature limits the search results when utilizing the same search query as used with Twint on the four biggest cryptocurrency subreddits. Because of this, Reddit could not provide enough data to perform sentiment analysis.

News stories and blogs seemed to have a similar issue when looking at hourly data. Not only is there no single place to easily access all published new stories and blogs in a specific time frame, but there also seems to not be enough content when using multiple different news APIs.

YouTube is another interesting social media platform where analyzing video titles could give enough data for sentiment analysis. The problem with this platform is that no API meets the requirements. The official YouTube API has limited functionality where every query has a maximum radius of 1000km of a location in the world, and this results in a bias that makes the platform unusable.

3.2.3 Price data

We need to get the closing price of a certain cryptocurrency every hour, so we can determine whether the price goes up or down in the next hour. We retrieve hourly price data by utilizing the Yahoo finance python package, which provides us with a comprehensive table of the Open, Close, High, Low, and trading volume for any COIN-USD pair that we require. This package retrieved all price data into a Pandas data frame, which can be easily modified. Here is an example of what the data looked like:

	Open	Close	High	Low	Volume
2022-04-01 01:00:00	46295.55	46242.56	46295.56	46238.35	1590110178
2022-04-01 02:00:00	46246.56	46249.92	46285.32	46213.01	1424680388
...					
2023-02-01 01:00:00	23370.68	23421.56	23391.81	23360.20	1602798302

Table 1: Price data of bitcoin

It is worth noting that the closing price often differed from the opening price of the subsequent hour, possibly due to trading activity occurring between the determination of the closing price and the new opening price.

3.3 Data ordering & cleaning

Since the social media data was obtained from Twitter, it often contains unstructured and cluttered information, requiring cleaning and structuring before performing natural language processing tasks. Twint provided the tweets along with all associated data in a CSV file. To organize the data properly and remove unnecessary information surrounding the tweets, we utilized the Pandas library. This allowed us to place the data in a more accessible format.

In addition to data ordering, cleaning was also essential. The sentiment analysis library we used operates on actual words and emojis. To address this, we removed non-alphabetic characters (excluding spaces) from the text.

For example, the following tweet:
"WOW... this is how it works... A perfect prediction so far... enjoy the pump which is coming ††^^^^^^

#Bitcoin #btc #BTC ”
Turns into:
”WOW this is how it works A perfect prediction so far enjoy the pump which is coming Bitcoin btc BTC ”
This step reduced a lot of unwanted characters in our data. It also removed punctuation, which is something we will discuss in Section 6.3. Furthermore, we implemented a feature in the script to eliminate duplicate tweets and mitigate the impact of Twitter's abundance of spambots. The link to all of the cleaned data can be found on the GitHub repository that belongs to this case study[1].

3.4 Target value

The target value of our machine learning model is to predict whether the closing price of the current hour, where the data is collected, will be higher than the closing price of the next hour. For each row of data, another entry is added with a binary variable relating to whether the price went up (True) or down (False). This target can be easily calculated using Table 1. The target is calculated in the same way for both sentiment and technical analysis.

[1] https://github.com/rutgerklaassen/Thesis-sentiment-analysishttps://github.com/rutgerklaassen/Thesis-sentiment-analysis

4 Methodology

This section provides the methods we used to create the models to generate our predictions.

4.1 Sentiment analysis

In this research, we utilize a lexicon-based approach by employing the VADER sentiment analysis tool [15]. VADER is used in this research for labelling the tweets with a sentiment score based on the words and sentence structure. VADER is a lexicon-based tool, which means that it uses a sentiment lexicon. A sentiment lexicon consists of lexical features that are categorized on their semantic orientation as positive or negative. Researchers typically construct a lexicon by compiling sentiment word lists using various approaches, including manual, lexical, and corpus-based methods. The polarity score of a given sentence or text is then determined by identifying the positive and negative indicators found in the lexicon. Next to giving a sentiment score to the words themselves, Hutto and Gilbert also performed qualitative analysis techniques to isolate five generalizable heuristics based on grammatical and semantic identifiers that could help determine the sentiment intensity. These heuristics are different from normal lexicon models as they incorporate word-order sensitive relationships between terms. The analysis interpreted the following five heuristics:

1. Capitalization, words that are capitalized have a stronger intensity than non-capitalized words. For example, "Bitcoin is a great investment" is less intense than "Bitcoin is a GREAT investment"
2. Degree modifiers, words that come before the sentiment-relevant word can impact the intensity. For example, "Bitcoin is an extremely good investment" is more intense than "Bitcoin is a good investment".
3. The contrastive conjunction, "but" usually signals a change in sentiment, with the sentiment of the text coming after the conjunction being dominant. "Bitcoin is a good investment, but it won't get you rich".
4. Negation, the negation of a sentiment-relevant word flips its polarity. For example, "Bitcoin isn't a good investment" is the opposite of "Bitcoin is a good investment"
5. Punctuation, specifically the exclamation mark can increase sentiment intensity without changing anything about the sentence itself. For example: "Bitcoin is a great investment" is less intense than "Bitcoin is a great investment!!!" This final heuristic was not properly applied due to our data sanitation. This conflict is addressed in Section 6.3.

The tweets in this research are analyzed based on their compound score. A compound score of a tweet is calculated by adding up all the sentiment scores of the individual or combination of words, which in turn is based on the lexicon and the above-mentioned heuristics.

After acquiring the compound score of a tweet, we used machine learning to try and predict price movements based on seven different parameters. The machine learning was performed using the Scitkit-learn [19] library. We used the following values as parameters for creating a machine-learning model:

- The average sentiment of an hour.
- The highest sentiment of the hour.
- The lowest sentiment of the hour.
- The number of tweets.
- The standard deviation of the sentiment scores.
- The skewness of the sentiment scores.
- The kurtosis of the sentiment scores.

The last three formulas were chosen due to them giving information about the distribution of sentiment, which could help the model in making its predictions.

4.2 Technical analysis

For performing technical analysis we used the technical indicators from the Pandas-ta library. With a requirement of only the price and/or volume data of select coins from up to two months prior, the library boasts more than 100 technical indicators. A technical indicator is a mathematical calculation based on historical market data like price, volume, and/or open interest of a security or contract used by traders to analyze price trends and make informed trading decisions. For our model, we utilized technical indicators from the following six groups: momentum, overlap, performance, trend, volatility, and volume [20]. All the technical indicators can be found on the Pandas-ta GitHub page under their respective categories. For example, we use the simple moving average technical indicator. The simple moving average refers to a technical indicator that calculates the average value of a set of prices over a specified period. This technical indicator is calculated by the following formula:

$$sma_n = \frac{P_1+P_2+...+P_n}{n}$$

Here $P_1 + P_2 + ... + P_n$ stands for the prices and n stands for the number of periods.
Another example of a technical indicator is momentum. This technical indicator is used to determine the difference between the last closing price and the closing price n days ago. The formula is as follows:

$$momentum_n = P - P_n$$

Here P is the last closing price and P_n was the closing price n periods ago where n has a default value of 10. We calculated all these technical indicators using cryptocurrency price data from up to two months before. After appending the target to it, we could train a model on it for predicting price movements.

4.3 Machine learning

From the Scikit library [19], we used the random forest [21] algorithm to create a classification model, which we used to make predictions on real-life data. The decision to employ the random forest algorithm in this thesis stems from its widespread adoption in the field, making it a normative choice. We also considered using autoML, as that algorithm will theoretically always have better results. This is due to the algorithm searching over every other machine-learning algorithm and checking which one performs the best, but this takes a lot of time. Because this research was time-limited, we decided not to use autoML. The code for the machine learning on both technical and sentiment analysis can be found on GitHub [2].

[2]https://github.com/rutgerklaassen/Thesis-sentiment-analysishttps://github.com/rutgerklaassen/Thesis-sentiment-analysis

5 Experimental setup

This section will explain which coins, preparation techniques, and evaluation methods we have used to generate our results

5.1 Cryptocurrencies

There are many coins available on the cryptocurrency market, with many of them having very different applications. We wanted to experiment on coins that had different purposes to prevent any bias that could occur due to the difference in applications. We opted to select more widely recognized coins, which allowed us to accumulate a sufficient number of tweets to train a model. We used the following four cryptocurrencies for our experiments:

- **Bitcoin**: The biggest and most well-known cryptocurrency [2].
- **Ethereum** : A programmable blockchain [22] and the second most popular coin [2].
- **Ripple**: An altcoin focused on creating a global currency for financial institutions [23].
- **Dogecoin**: The first altcoin focused on community-building around a crypto-asset, primarily through social media.

5.2 Baseline

It is important to compare our models to a baseline model. This way we can see if our models are an actual improvement. For this, we used the majority class classifier. This creates a baseline for us to use and compare our models to. The model determines what the most frequently occurring class is in the training data, and predicts that for every entry in the test data. For instance, when training a model to predict cryptocurrency price movements, if 53% of the cases in the training set involve upward price movements, the model will consistently predict upward price movements for the test set.

5.3 Accuracy

To determine whether a model performs well or not, we selected accuracy as an evaluation measure. The accuracy of all models was determined by summing up the accurate predictions of price increase and decrease, and then dividing this sum by the total number of predictions.

5.4 Evaluation protocol

For this thesis, we gathered 275 consecutive days of data to train our models. The training set was formed with data from 2022-04-01 until 2023-01-01. The minimal amount of training data was set to 10 days of data, and the maximum was set to 275. After training a model, we added the next 5 days of data and trained a new model. This summed up to a total of 54 models per coin, ranging in training data from 10 days up to 275 days. Following the completion of model training, we proceeded to evaluate the performance of a model over a subsequent period of 30 days. This means that if a model was trained on the first 10 days of training data, it was then tested on the 30 days that follow the 10 days it was trained on. For a schematic overview of our evaluation protocol, see Figure 2.

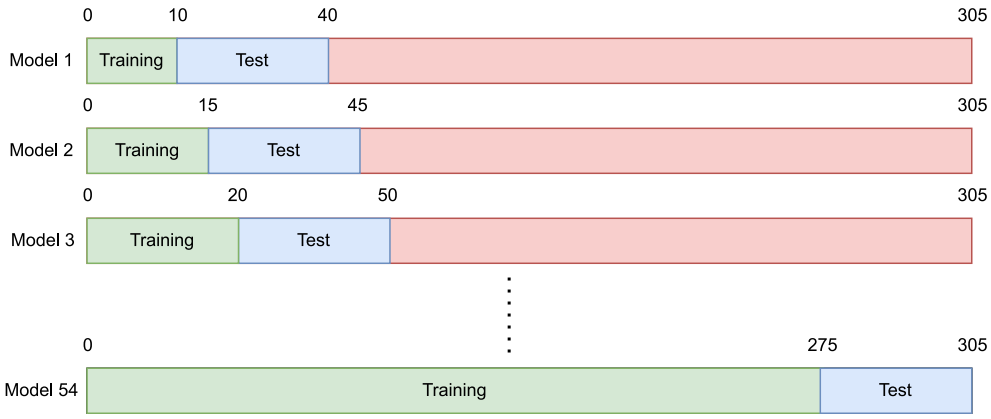

Figure 2: A visual representation of the evaluation protocol used in this research

6 Results

In this section, we will discuss the results we gathered in the experiments from section 5. We will show the accuracy scores of the different models that are trained on varying data sources, and how it compares to the baseline (majority class classifier).

6.1 General trend

This subsection will show and explain different perspectives on our results.

Average accuracy.
The following table displays the average accuracy and standard deviation (inside the brackets) for each prediction in the test set.

Coin	Baseline	Sentiment	Technical	Combination
Bitcoin	49.76 (±1.97)	50.29 (±1.51)	51.56 (±1.82)	51.62 (±1.85)
Ethereum	50.34 (±1.67)	49.84 (±1.83)	50.97 (±1.92)	50.79 (±1.60)
Doge	50.36 (±1.58)	49.87 (±1.70)	51.80 (±1.71)	51.92 (±1.53)
Ripple	51.17 (±2.07)	49.84 (±1.57)	52.10 (±1.50)	51.51 (±1.54)

Table 2: Average accuracy scores and standard deviation calculated from models with training times varying from 10 to 275 days

We can see that on average the models trained on sentiment data do not outperform the majority class classifier. The model trained on Bitcoin is also the only model where the sentiment analysis performed (slightly) better than the majority class classifier. Another interesting observation is that training a model on sentiment data performs worse than training it on technical data. The models trained on technical data consistently perform better than both the majority class classifier and the models trained on sentiment data. Furthermore, there are instances where the incorporation of sentiment data negatively impacts prediction accuracy as evidenced by the combination of technical and sentiment data, specifically when examining Ripple and Ethereum.

Accuracy compared to the amount days trained.
In Figures 2 - 9, we can see how the model performs with the increasing size of the training. The length of the test set that is used to measure the accuracy stays the same (30 days). Every coin has two graphs, one graph showcasing the results of all the models, and one graph that has the trendlines for all the different types of data. In both graphs, the blue lines stand for the baseline, which is the majority class classifier.

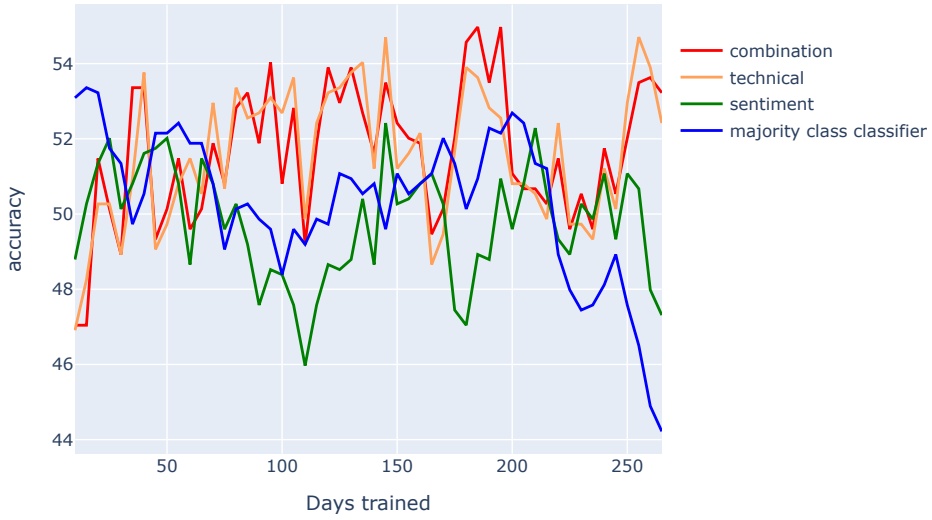

Figure 2: Comparison of the accuracy of a random forest model trained on Bitcoin sentiment data, technical data, and their combination, relative to a majority class classifier, across varying training durations

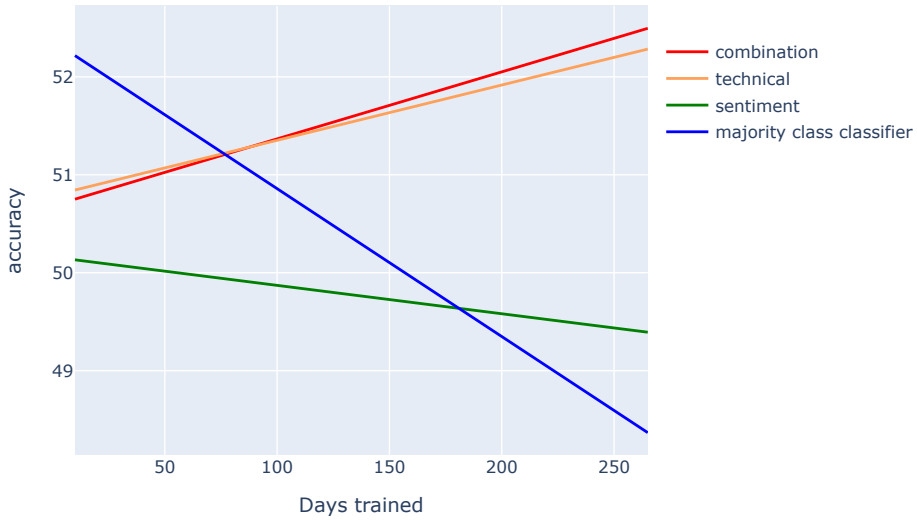

Figure 3: Trendline of all random forest models trained on Bitcoin sentiment data, technical data, and their combination, relative to a majority class classifier

Figure 4: Comparison of the accuracy of a random forest model trained on Dogecoin sentiment data, technical data, and their combination, relative to a majority class classifier, across varying training durations

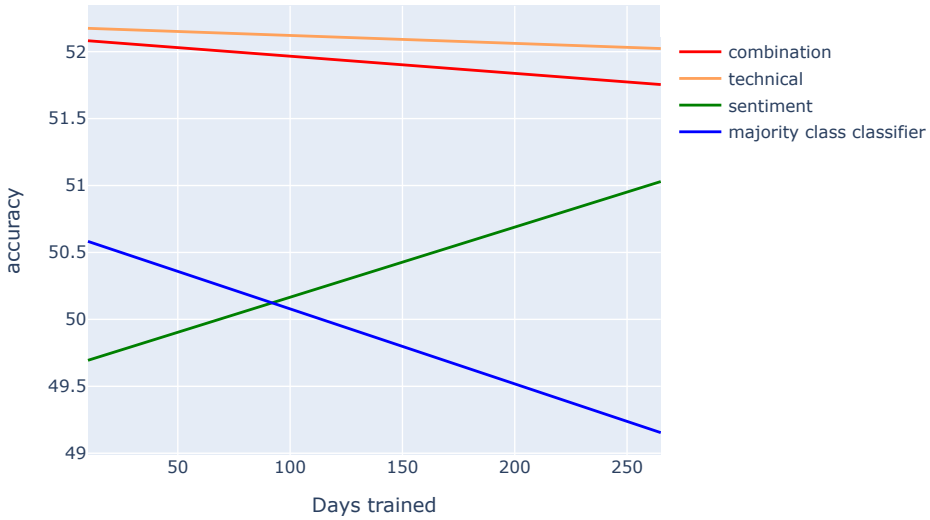

Figure 5: Trendline of all random forest models trained on Dogecoin sentiment data, technical data, and their combination, relative to a majority class classifier

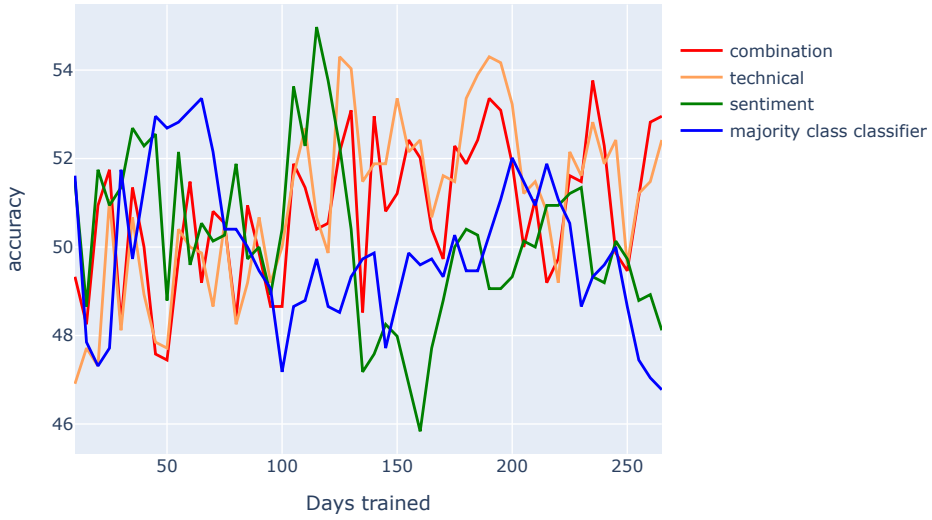

Figure 6: Comparison of the accuracy of a random forest model trained on Ethereum sentiment data, technical data, and their combination, relative to a majority class classifier, across varying training durations

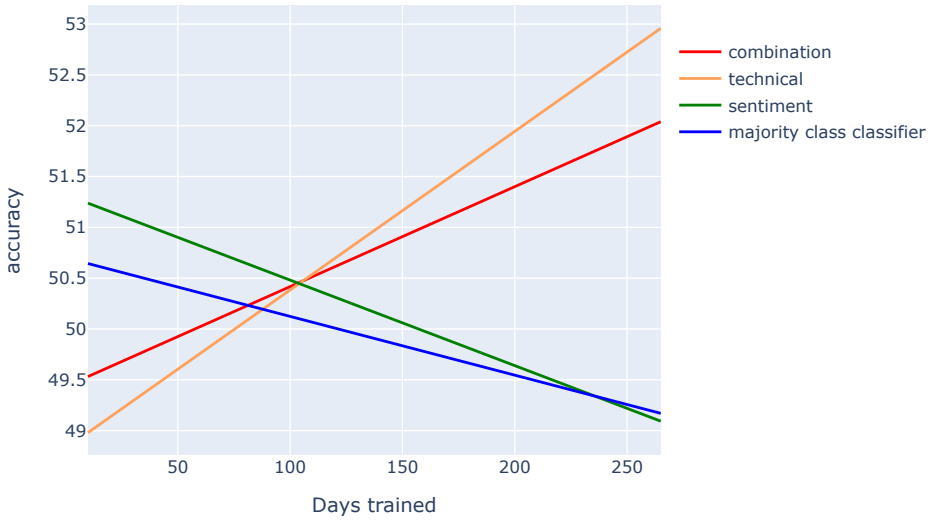

Figure 7: Trendline of all random forest models trained on Ethereum sentiment data, technical data, and their combination, relative to a majority class classifier

Figure 8: Comparison of the accuracy of a random forest model trained on Ripple sentiment data, technical data, and their combination, relative to a majority class classifier, across varying training durations

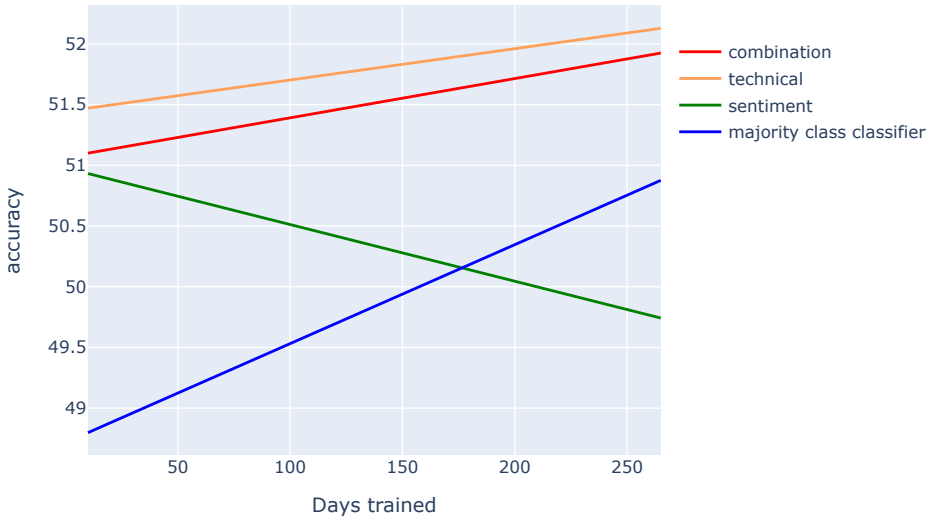

Figure 9: Trendline of all random forest models trained on Ripple sentiment data, technical data, and their combination, relative to a majority class classifier

We can see from Figures 3,5,7, and 9 that increasing the training time makes the model better on average for technical analysis and the combination of technical and sentiment analysis for Bitcoin, Ethereum, and Ripple. For these three coins, the models trained on sentiment analysis alone do not seem to gain better accuracy with longer training times. It is interesting to notice that the opposite seems to be the case for Dogecoin, where the accuracy of models trained on technical analysis and the combination of technical and sentiment analysis decreases as the training time increases and the accuracy of the models trained on just sentiment data increases as the training time gets longer.

Another noteworthy observation is that whenever the accuracy of the majority class classifier descends towards the lower boundary of its range, around 48%, the models based on technical analysis and the combined approach tend to excel. This is noticeable in Figure 2 around day 250, in Figure 4 around day 100 and day 150, in Figure 6 around day 250 and in Figure 8 around day 100 and around day 175. The opposite is also noticeable in Figure 2 around day 200, in Figure 4 around day 160, in Figure 6 around day 50, and in Figure 8 around day 125 and day 240.

Boxplot of accuracies In Figure 10 we show the spread of accuracies in a boxplot graph. It is important to notice that the line in the middle of the boxes is the median, not the mean, which is why they're different from the values in Table 2.

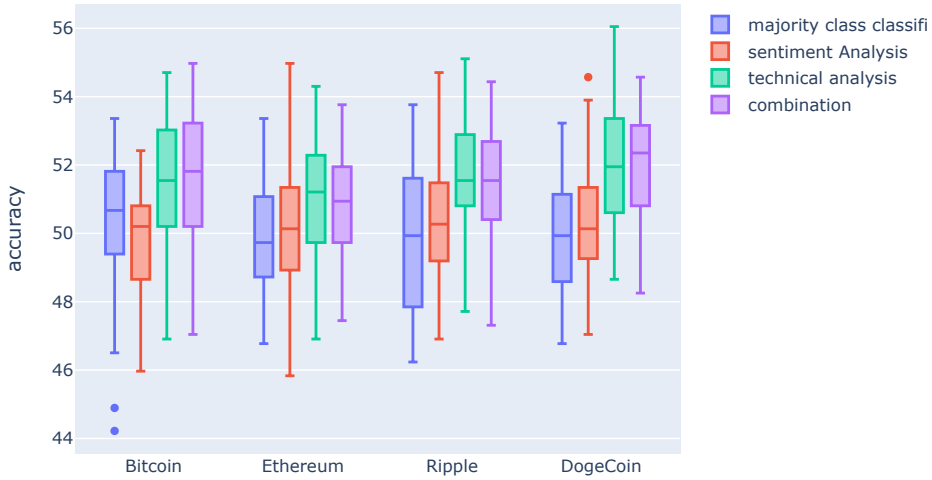

Figure 10: Boxplot of accuracies from a random forest model trained on sentiment data, technical data, and the combination of them both compared to the majority class classifier for four different coins

Looking at these boxplots, we can observe a difference between these plots and Table 2. We can see that the models based on technical data and the combination of technical and sentiment data consistently perform better than the majority class classifier, which is the same. Another interesting observation is that we do see that the models trained on sentiment data slightly outperform the majority class classifier. So even though the mean of the models trained on sentiment data is lower than the majority class classifier, the median of these is slightly higher.

6.2 Statistical analysis

The plots from the previous subsection show that certain models outperform the baseline. To evaluate whether the obtained results appear significant from a statistical viewpoint, a statistical significance test should be performed. Based on the suggestions by Demšar [24], we performed the Friedman test ($\alpha = 0.05$) with the Nemenyi method as a post-hoc test [25]. The Friedman test functions as a way to reject the null hypothesis, which is that the performance of all the classifiers is equal. The Nemenyi-test shows whether performance differences between models are statistically significant. The test is based upon differences in classifier ranking, such a ranking is obtained by comparing the performance of different classifiers on the same data set, and by comparing the models on all 52 datasets we can compute an average rank.
Table 3 shows the results of the Friedman test, Figures 11-14 show the results of the Nemenyi-test

Coin	p-value
Bitcoin	1.86e-05
Ethereum	0.015
Dogecoin	9.05e-08
Ripple	8.52e-08

Table 3: p-value of the of the different classifier for each coin

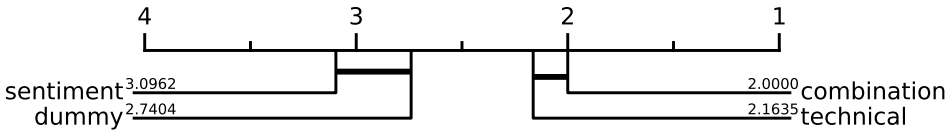

Figure 11: Comparison of all Bitcoin classifiers against each other with the Nemenyi-test

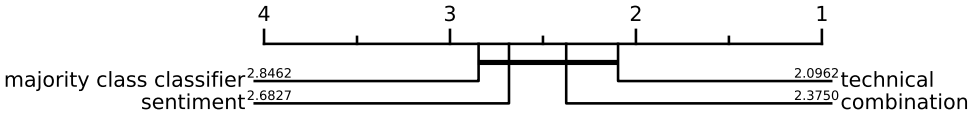

Figure 12: Comparison of all Ethereum classifiers against each other with the Nemenyi-test

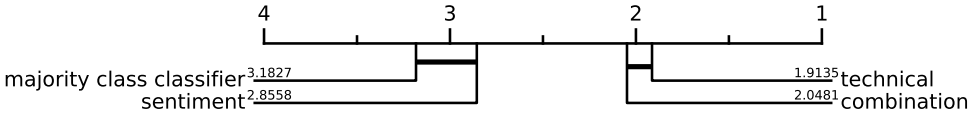

Figure 13: Comparison of all Doge classifiers against each other with the Nemenyi-test

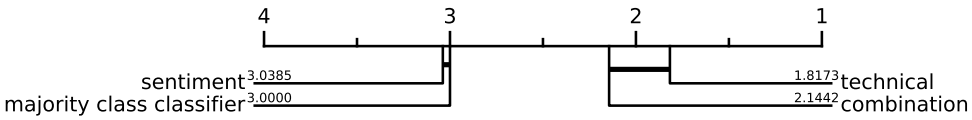

Figure 14: Comparison of all Ripple classifiers against each other with the Nemenyi-test

In Table 3 we can see that there is a statistically significant difference between the classifiers for every coin, as every value is below 0.05. From the Nemenyi-tests, a significant difference can be observed between the majority class classifier and the technical analysis, and the majority classifier and the combination of technical and sentiment analysis. This is true for each model except for the classifier predicting the price trend of Ethereum. The Nemenyi-test also shows that the performance of the models trained on sentiment analysis is not statistically significant compared to the majority class classifier.

6.3 The impact of punctuation

For this research, we used datasets gathered from Twitter and sanitized these datasets to make sure the sentiment analysis tool would only get alphanumeric characters. The sanitation process also removed punctuation, a significant heuristic that constitutes a vital component of sentiment analysis. This subsection serves as an analytical exploration of how the presence or absence of punctuation affects the outcomes of this study. More specifically, we will focus on the data set of Bitcoin, comparing the data set without punctuation to the data set with punctuation. In Figures 15-18, the results of these comparisons are shown.

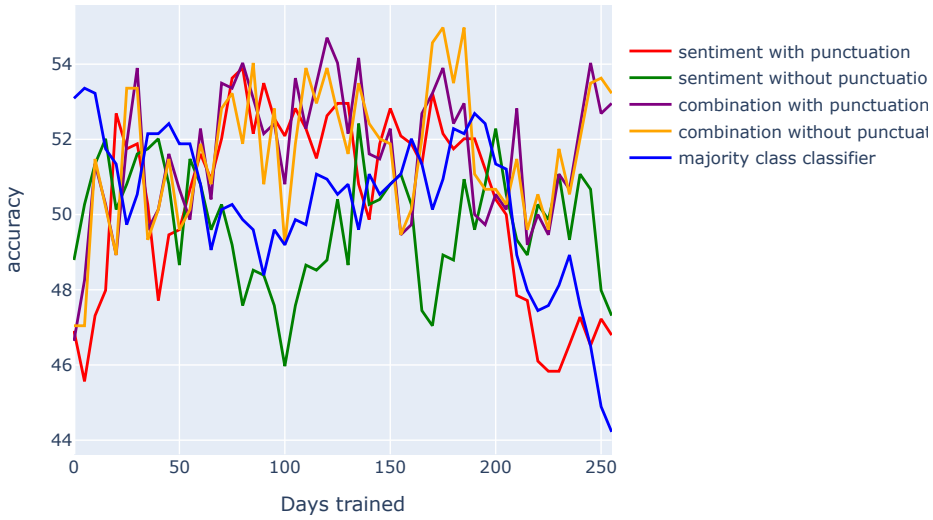

Figure 15: Comparison of sentiment models trained on Bitcoin data with and without punctuation relative to the majority class classifier

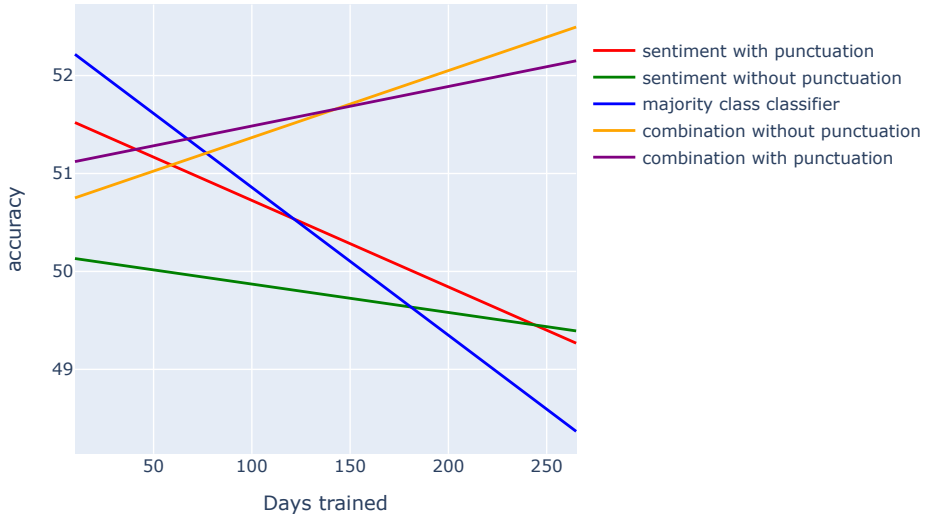

Figure 16: Trendlines of combination models trained on Bitcoin data with and without punctuation relative to the majority class classifier

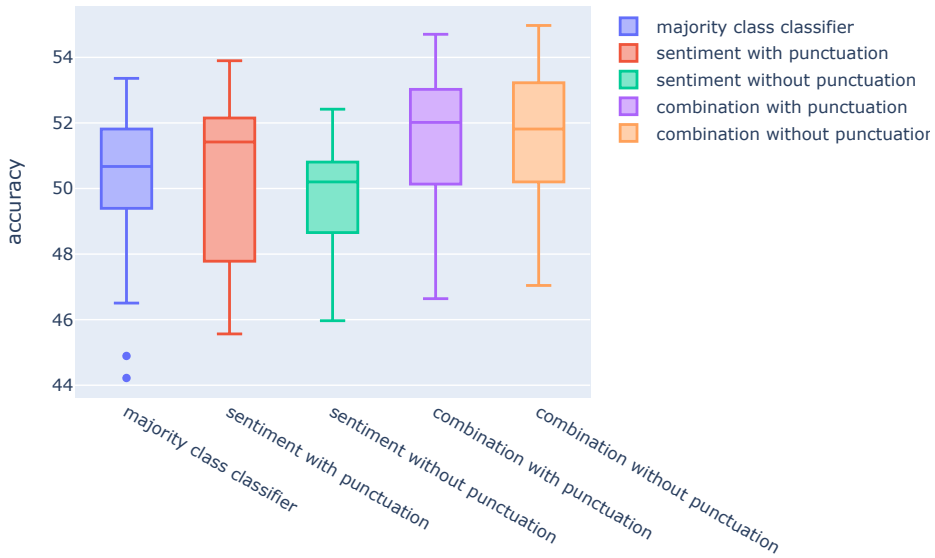

Figure 17: Boxplots of models trained on Bitcoin data with and without punctuation relative to the majority class classifier

In these figures, we observe that the models trained on data with punctuation outperform the models without punctuation for sentiment analysis. Punctuation makes a noticeable difference, as the figures show that the sentiment model trained on data with punctuation marks outperforms the majority class classifier, whereas the sentiment model trained on data without punctuation does not. The same

cannot be said for the combination of technical and sentiment analysis, as their performance is nearly identical. Performing a statistical significance test also shows an improvement in average ranking, as can be seen in Figure 18. However, it does not constitute a significant difference between the majority class classifier and the model trained on sentiment analysis with punctuation.

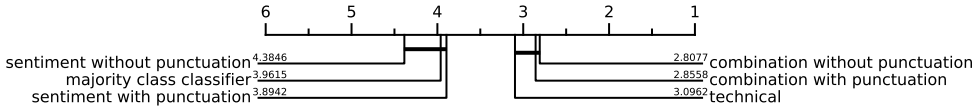

Figure 18: Comparison of all Bitcoin classifiers against each other with the Nemenyi-test

From this, we can conclude that while sentiment analysis works better with punctuation for Bitcoin, it still does not have a significantly different performance compared to the majority class classifier. We have also performed the Friedman test on all sentiment classifiers to see if the model trained on sentiment analysis with punctuation is significantly different from the models trained on sentiment analysis without punctuation. The Friedman test resulted in an α that was greater than 0.05, which means that we cannot reject the null hypothesis that the performance of these models is equal.

7 Discussion & Limitations

This study employed Twint to gather Twitter data due to limited access to the Twitter API. It is essential to acknowledge that this approach may not capture the entirety of tweets posted during the requested time frame. Twint's functionality and coverage might differ from the official API, potentially leading to incomplete data collection. Consequently, the sample of tweets analyzed may not represent the entire spectrum of the sentiment on Twitter, thereby introducing a potential bias in the results.

Another limitation stems from the reliance on Twitter as the only social media source. While Twitter is a popular platform for cryptocurrency discussions, it is crucial to recognize that the sentiment on other social media platforms may differ, which means that by only using Twitter for sentiment analysis we may introduce a bias. The findings related to sentiment analysis should be interpreted within the context of Twitter users' opinions and may not generalize to sentiments expressed on the entire internet.

Despite our efforts to sanitize all the data, it is possible that some tweets from bots were not removed from the data set. A lot of these bots repost their tweets with only a few different characters which evades our detection and potentially influences our sentiment analysis results.

Furthermore, this study focused on utilizing a specific set of machine-learning models for prediction purposes. It is worth noting that different models may yield varying results. Previous research has indicated that random forest models might not perform as well as other models for sentiment analysis tasks. Therefore, our results showing that sentiment analysis does not perform better than randomly guessing may be due to the limited effectiveness of the chosen machine learning algorithm on sentiment analysis.

This research used a one-hour interval to gather input data for creating machine learning models. This is because the one-hour interval serves as the median between the very short durations (1 minute, 5 minutes, 15 minutes) and the longer ones (4 hours, 12 hours, up to a full day). It is possible that this limited our results, as algorithmic trading of cryptocurrencies has shown promise in shorter intervals [12].

It is also possible that worldwide events could alter price movements, which could potentially reduce the effectiveness of model training by introducing volatile changes to the price of a cryptocurrency.

Another factor to consider is that the training times might be too small. Twint is a very unoptimized library and gathers tweets by entering queries into a Twitter search bar. This slows down the process of gathering tweets massively compared to using the official Twitter API. Next to that, after running Twint for more than half a day, the API seems to put a limit on how fast you can gather the tweets, which slows down the process even further. The precise origin of this limitation, whether from the API or Twitter itself, remains uncertain. However, it renders the collection of a larger volume of tweets within the time frame allocated for this study infeasible. Due to this, we could not allow more training data, which might have led to improved results. Another implication of the limited training data is the time frame of the training. All of the tweets were collected in a time frame of 9 months. It is possible that the shorter time frame only catches a market that is in an either upward or downward trend, making the model possibly less effective when the market is in the opposite trend in the test set.

Possibly the biggest limitation of this study is the error relating to punctuation. By removing the punctuation from our data, we limited the performance of our sentiment analysis to the point where it was worse than the majority class classifier. Even though a Nemenyi test showed there was no statistically significant difference between the models trained with and without punctuation, we can still assume based on Figures 15-17 that our sentiment analysis models would have performed better with punctuation.

It is important to consider that this study is part of a larger pipeline, where synergy with the other parts is important. When applying the techniques outlined in this study to develop a trading strategy, it is vital to be aware that challenges may emerge if the components don't complement each other effectively. The study itself is a pipeline as well, so if the data gathering and cleaning parts do not work well with the machine learning process it could lead to inadequate results. It is also important to note that due to the lack of backtesting it is unknown whether a profitable trading strategy can be developed from the methods described in this research.

Lastly, the accuracy calculations were initiated using models trained with as few as 10 days' worth of tweets. This minimum training time might have a negative impact, and lower the accuracy, whereas if we started with a way higher training time we might have seen a higher average accuracy. The reason for having a shorter training time is the lack of data. If we had more time to gather data, we could implement a higher minimum training time, and possibly increase the average accuracy of our models.

8 Conclusion

This research is a case study on sentiment analysis within Twitter aimed to investigate the effectiveness of sentiment and technical analysis in predicting the price movements of four different cryptocurrencies. A binary classification task was introduced to investigate the potential increase or decrease in the valuation of a cryptocurrency within the upcoming hour. We used machine learning on both technical and sentiment analysis to create models and predict the target.

The machine learning models based on technical analysis were trained on technical indicators that were calculated based on the open, close, high, low, and volume attributes of the coin. The machine learning models based on sentiment analysis were trained on sentiment scores that were gathered from using the VADER sentiment analysis tool on tweets relating to a specific cryptocurrency. This thesis also attempted to investigate the comparative effectiveness of sentiment data from specific social media sources; however, it encountered challenges in gathering sufficient training data for sentiment analysis due to limited publicly available APIs.

Models were trained with varying amounts of training data as an evaluation protocol. To see whether our models performed well, we established a baseline model with the majority class classifier to compare our models to. The experiments carried out and the subsequent analysis yielded valuable insights regarding the performance of various models, along with their comparison to a baseline.

Analyzing the results, it was found that sentiment analysis alone did not yield superior results compared to the majority class classifier for all coins when trained on data without punctuation. The boxplot analysis shows us something slightly different. By examining the median of all the models, sentiment analysis without punctuation did seem to perform slightly better than the majority class classifier. It appears that the model trained on sentiment data including punctuation exhibited superior performance compared to models trained without punctuation. However, we did not discover statistical evidence to support this observation. The model trained on a combination of sentiment and technical data did not benefit from adding punctuation. Furthermore, training models on technical data generally outperformed those trained on sentiment data. By performing statistical significance tests we found statistical evidence that models trained on technical analysis as well as models trained on the combination of technical and sentiment analysis perform better than both the majority class classifier and sentiment analysis. Surprisingly, the inclusion of sentiment data in technical analysis occasionally led to decreased prediction accuracy. The statistical significance tests did not find any evidence that there was a statistical difference between technical analysis and the combination of technical and sentiment analysis.

Examining the impact of training time on model performance, it was generally observed that longer training periods improved the accuracy of models based on technical analysis and the combined approach for models trained on Bitcoin, Ethereum, and Ripple. However, for models trained on Dogecoin, accuracy seemed to decrease with a longer training time. The accuracy of sentiment-based models did not consistently increase with more training data, except for Dogecoin, where it did consistently perform better with an increase in training data.

With these results, we can try to answer our research question: *Is there a correlation between media sentiment analysis and cryptocurrency price movements, and how does it compare to technical analysis?*. According to our results, there is a slight correlation between media sentiment analysis and cryptocurrency price movements. Although the superiority of sentiment analysis models might not be immediately apparent from the average accuracies of all the models, and despite the absence of statistical evidence, the median performance of our sentiment analysis models consistently surpasses that of our baseline. We can also conclude that in comparison to technical analysis, sentiment analysis is less effective in predicting cryptocurrency price movement. Overall, this research provides valuable insights into the effectiveness of technical analysis as well as sentiment analysis in cryptocurrency price movement prediction. As a case study, this research added evidence proving the effectiveness of mainly technical analysis. The findings suggest that while sentiment analysis using a random forest model alone may not offer significant advantages over a majority class classifier, random forest models incorporating technical analysis show more promise in capturing market dynamics.

9 Further research

Future research on this topic could explore alternative methods for data collection by utilizing other social media platforms. Multiple platforms have a big enough audience to make an impact on the sentiment score of a specific cryptocurrency. Reddit, Facebook, financial news articles, and YouTube are all options researchers should consider when performing sentiment analysis. This would help gather a more comprehensive and diverse data set, enabling a broader analysis of sentiment across multiple platforms and potentially reducing biases associated with relying solely on Twitter. When obtaining data, future researchers should make sure to include punctuation marks, as sentiment analysis seems to perform slightly better when they're included.

Furthermore, future researchers could investigate the use of training data from different time frames as it might give better signals and increase the accuracies.

Next to that, many promising machine-learning techniques seem to perform sentiment analysis with a higher accuracy [6] [16], such as multi-layer perceptrons, support vector machines, and the naive-Bayes classifier.

In our machine-learning models, there is room for improvement in the form of hyperparameter optimization. Studies have shown that hyperparameter optimization can increase the accuracy of random forest models [26, 27]. This could be advantageous as it would result in higher accuracy for our sentiment analysis as well as our technical analysis models.

Further studies can also make these findings applicable by developing a trading strategy with it. We encourage other researchers to improve and use this work in their research.

References

[1] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, page 21260, 2008.

[2] Coinmarketcap. `https://coinmarketcap.com/all/views/all/`. Accessed: 2023-04-06.

[3] Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions. *Electronics*, 10(21):2717, 2021.

[4] Shaen Corbet, Veysel Eraslan, Brian Lucey, and Ahmet Sensoy. The effectiveness of technical trading rules in cryptocurrency markets. *Finance Research Letters*, 31:32–37, 2019.

[5] M Poongodi, Tu N Nguyen, Mounir Hamdi, and Korhan Cengiz. Global cryptocurrency trend prediction using social media. *Information Processing & Management*, 58(6):102708, 2021.

[6] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6):589, 2019.

[7] Wei Yen Chong, Bhawani Selvaretnam, and Lay-Ki Soon. Natural language processing for sentiment analysis: an exploratory analysis on tweets. In *2014 4th international conference on artificial intelligence with applications in engineering and technology*, pages 212–217. IEEE, 2014.

[8] Connor Lamon, Eric Nielsen, and Eric Redondo. Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev*, 1(3):1–22, 2017.

[9] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. In *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*, pages 339–343. IEEE, 2018.

[10] Mohammed Mudassir, Shada Bennbaia, Devrim Unal, and Mohammad Hammoudeh. Time-series forecasting of bitcoin prices using high-dimensional features: a machine learning approach. *Neural computing and applications*, pages 1–15, 2020.

[11] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[12] Stan van der Avoird. Prediction and technical analysis of the bitcoin crypto currency using machine learning. 2020.

[13] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[14] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[15] Clayton Hutto and Eric Gilbert. 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*, volume 8, pages 216–225, 2014.

[16] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. *CS229 Project*, pages 1–5, 2015.

[17] Twint: Twitter intelligence tool. https://github.com/twintproject/twint.

[18] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Kevin Johnson. Pandas-ta. `https://github.com/twopirllc/pandas-ta`.

[21] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[22] Gavin Wood et al. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151(2014):1–32, 2014.

[23] David Schwartz, Noah Youngs, Arthur Britto, et al. The ripple protocol consensus algorithm. *Ripple Labs Inc White Paper*, 5(8):151, 2014.

[24] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[25] Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

[26] Fabian Schut. Machine learning and technical analysis for foreign exchange data with automated trading. 2019.

[27] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.