TITLE: INSTRUCTIONS FOR THE AUTHORS OF PAPERS

Urs A. Hurni

University of Lausanne

Lausanne, Switzerland

urs.hurni@unil.ch

1 ABSTRACT

This research examines the influence of market sentiment on cryptocurrency prices through sentiment analysis and machine learning techniques, utilizing data from news articles and Bitcoin price trends. Sentiment is mesured using the FinBERT model. The study involves several steps, like data resampling, various techniques such as ARIMA and Random Forest. Additionally, the research conducts correlation analysis to explore the relationship between sentiment scores and price shifts, using lag features to optimize prediction timings. The findings suggest a moderate, often fluctuating connection between market sentiment and price changes and short-term prediction proving to me more accurate, highlighting the complexity of market reactions to news. The random forest coupled with bootstrapping proved to be the most effective model. The study underscores the challenges and potential of leveraging sentiment analysis for cryptocurrency predictions, emphasizing the necessity of sophisticated data handling and advanced modelling techniques to enhance prediction accuracy.

2 INTRODUCTION

Advances in machine learning and natural language processing (NLP) have transformed how data is utilized, offering new perspectives on market trends and potentially enhancing predictive models for crypto price movements. As we delve deeper into the crypto currency special universe, the role of sentiment analysis becomes increasingly crucial. By using sophisticated algorithms and machine learning techniques, analysts and investors are now able to decrypt vast amounts of unstructured data from news articles, social media, and financial reports to gauge public sentiment toward various financial instruments, including cryptocurrencies.

Cryptocurrencies, by their very nature, are highly volatile and influenced by a wide array of factors ranging from global economic indicators to regulatory news and even social media trends. Traditional financial models, while still relevant, often fall short when tasked with capturing the swift shifts in investor sentiment that can drastically affect crypto markets. This is where AI and NLP stand out, providing the tools necessary to analyze and interpret the mood and opinions of the market at large, translating this data into actionable insights that can precede market movements

The integration of sentiment analysis into financial decision-making processes marks a significant shift towards data-driven strategies. This approach not only enhances the understanding of market dynamics but also aids in the development of more robust trading systems that can better withstand the unpredictability of the crypto markets. By applying sentiment analysis, traders can identify potential buying or selling signals based on the collective emotions of market participants, thereby gaining a strategic edge.

Moreover, the ability of NLP to process and analyze real-time data allows for a more agile response to market changes. In the fast-paced world of cryptocurrency trading, where prices can fluctuate wildly within minutes, the speed at which data is processed and interpreted is crucial.

However, the application forecasting, machine learning and NLP in sentiment analysis is not without its challenges. One of the primary concerns is the accuracy of the sentiment gauged from various sources. The contextual nuances of language, sarcasm, and misleading information can sometimes skew the analysis, leading to potentially erroneous conclusions. Furthermore, the sentiment itself is highly subjective and can be influenced by temporary external factors that do not necessarily reflect the long-term market trends.

Despite these challenges, the potential benefits of integrating sentiment analysis into cryptocurrency trading and broader financial strategies are significant. It opens up new avenues for research and development within financial technology, encouraging further innovation and refinement of analytical tools

3 RESEARCH QUESTION

How effectively can market sentiment derived from various sources predict cryptocurrency price movements in the short-term and long-term, and what is the optimal lag time between sentiment shifts and price changes for accurate predictions?

3.1 Problem

The main issue being addressed is the effect of market sentiment on cryptocurrency prices and how to enhance the prediction of these price movements through sentiment analysis. Cryptocurrencies are notably volatile and highly sensitive to public sentiment, which is quickly spread via social media, news outlets, and other digital channels. The challenge lies in quantitatively evaluating how changes in public mood and opinion influence cryptocurrency values over both short and long terms.

3.2 Objective

The objective is to quantify the impact of market sentiment on cryptocurrency prices and determine if sentiment alone can predict short-term and long-term price movements. The primary objectives of this project are:

- To quantify the impact of market sentiment on cryptocurrency prices and identify if sentiment only can predict short-term and long-term price movements.
- To develop a predictive model that utilizes sentiment scores from various sources to forecast cryptocurrency price changes
- To examine the lag effect between sentiment shifts and price changes, determining optimal time frames for predicting price movements based on sentiment analysis.

3.3 Literature Review

In this section, various studies that have explored the use of sentiment analysis for financial forecasting are reviewed, highlighting the methods and findings relevant to our research.

Fazlija and Harder [3] explored the use of sentiment from financial news articles to forecast the direction of the S&P 500 index. They applied natural language processing (NLP) to determine sentiment scores from articles and incorporated these scores into predictive models like the random forest classifier. Their findings suggest that news-based sentiment scores are effective in predicting stock movements, underscoring the value of sentiment analysis in financial forecasting.

Souma, Vodenska, and Aoyama [5] investigated the potential of deep learning, particularly transformer models such as BERT, for sentiment analysis in financial texts. Their results showed that these advanced models could accurately gauge sentiment and significantly improve stock price prediction accuracy, suggesting their possible application to cryptocurrencies.

T. Adams and all [4] conducted a study on using Twitter sentiment analysis to predict stock market trends. They concluded that sentiments expressed on social media could be powerful indicators for market movements. Given the active discussion about cryptocurrencies on platforms like Twitter, these methods might also predict trends in cryptocurrency prices effectively.

M. P. Cristescu [6] focused on how news sentiment impacts stock prices, combining sentiment analysis with traditional financial metrics. They discovered that adding sentiment scores to past price data considerably enhances the accuracy of predictions. This method could be particularly useful in the cryptocurrency sector, where market sentiment frequently influences price changes.

Research [7] has also highlighted the effectiveness of random forest classifiers in forecasting stock prices using sentiment scores. By combining several decision trees, this method improves prediction accuracy and could be adapted to analyze cryptocurrency market trends, utilizing sentiment data for better forecasting accuracy.

Despite the potential benefits, sentiment analysis in financial markets faces several challenges. The accuracy of sentiment detection can be affected by the contextual nuances of language, sarcasm, and misleading information. Additionally, sentiment is inherently subjective and can be influenced by transient external factors that do not reflect long-term market trends. These challenges must be addressed to improve the reliability of sentiment-based predictive models [8]

3.4 Scope

For the further analysis we are going to use a set of a pre-loaded dataset.

Data to be used are :

- 2024-05-18_24h_news_with_sentiment.csv
- 2024-05-18_24h_50meme_history.csv
- 2024-04-05_30d_news_with_sentiment.csv
- 2024-04-05_30d_50meme_history.csv
- Daily Bitcoin Price.csv
- testing_news_with_sentiment_data.csv
- testing_price_data.csv
- Bitcoin Search Trend.csv

These 4 first datasets are representative of the type of data typically accessible through the APIs, but to adhere to the constraints of the free API tier and minimize data fetching, we are using these static files. While the results presented are based on this specific data, the methodology developed is designed to be adaptable and replicable across different sectors.

The other files are used for historical analysis.

4 METHODOLOGY

The methodology section of this research outlines the systematic steps and processes used to explore the influence of market sentiment on cryptocurrency prices.

4.1 Data Processing

Processing was needed as the news often came from various sources and were not all on the same format.

Indeed, the news data needed to be cleaned to include the date, headline, and text for each entry. This is necessary for it to be processed by the sentiment function, which assigns a sentiment score ranging from 0 to 1 to each news item. The price dataset also required normalization to a scale between 0 and 1. This step is crucial for merging it with the sentiment data for further analysis.

The combined dataset must contain columns for date, average sentiment, price, and price change. It should be indexed by date to enable accurate time-based resampling and alignment. Proper column naming is essential for the analysis function to process the data effectively. This setup ensures that the models can work correctly with the prepared data.

Further detail on this is explored in the 'Dataset' section of this report.

4.2 Sentiment

To start analyzing sentiment trends, a Google Trends analysis on Bitcoin prices was conducted to see how search volume correlates with cryptocurrency price movements. This basic analysis sets the foundation for getting a sense on how a simple metrics influenced price. Google Trends provides an estimate of search volume, which helps us explore if search popularity relates to other types of data. Patterns were searched in the Google search volume and Bitcoin prices to see if there is a connection.

Next, we used a pre-trained model called FinBERT to define sentiment scores. FinBERT is based on BERT (Bidirectional Encoder Representations from Transformers) and is further trained on financial texts. This extra training helps FinBERT understand the unique vocabulary and expressions used in financial articles [9]. The model is provided by ProsusAI and is fine-tuned to interpret the nuances of financial language. This makes it especially useful for analyzing sentiment in news articles related to cryptocurrency.

Using FinBERT, the sentiment embedded in cryptocurrency-related news can be assessed. This helps understand how news sentiment might influence Bitcoin prices. By combining the insights from Google Trends and FinBERT, we can have a better understanding of the relationship between public interest, news sentiment, and cryptocurrency price movements.

4.3 Model Application and Description

4.3.1 ARIMA

The ARIMA (AutoRegressive Integrated Moving Average) model is widely used for predicting future price changes in cryptocurrencies, leveraging historical price data. This model is particularly effective in scenarios where the data shows trends or seasonal patterns. In cases where such patterns are absent, ARIMA can switch to a simpler, naive approach, typically using the most recent observed value as the forecast. The *'auto'* aspect of ARIMA is designed to adapt to the incoming data by automatically selecting the best parameters (p, d, q) to optimize the model. This optimization aims to minimize forecasting errors, guided by criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). ARIMA then predicts the average percentage change in future cryptocurrency prices over a predetermined number of periods.

4.3.2 Correlation

In addition, a correlation analysis measures the strength and direction of the relationship between sentiment scores and subsequent price changes. Pearson correlation coefficients is used to pinpoint the strongest predictive relationships. This method helps in quantifying how sentiment, as expressed through various metrics, aligns with and potentially influences future price changes.

4.3.3 Linear Regression

Linear regression in this context utilizes the Ordinary Least Squares (*OLS*) method to analyse the predictive power of sentiment on cryptocurrency price. This model assumes a linear relationship between the independent variable (*sentiment scores*) and the dependent variable (*price changes*). The strength of this approach lies in its straightforwardness and the interpretability of its results, which indicate how sentiment scores quantitatively impact price changes in the market.

4.3.4 Random Forest

The Random Forest model employs the *RandomForestRegressor* for predicting cryptocurrency price changes using sentiment data. This approach is non-linear, allowing it to model more complex relationships than linear models. Random Forest uses an ensemble learning technique, which involves the construction of multiple decision trees during training and outputs the mean prediction of the individual trees. This method enhances the robustness and accuracy of predictions by

reducing the risk of overfitting to the training data and improving performance on unseen data.

4.3.5 Support Vector Regression

Support Vector Regression (SVR) is used to predict continuous outcomes like changes in cryptocurrency prices. SVR works by finding the best-fitting line within a set margin, making it good for handling non-linear relationships. It uses kernel functions to transform the data into a higher-dimensional space, which helps in managing complex patterns that a simple linear model might miss. One of the main benefits of SVR is its ability to balance model complexity with prediction accuracy, reducing errors effectively. This method is especially helpful when the relationship between sentiment data and price changes isn't straightforward.

4.3.6 Gradient Boosting Machine

The Gradient Boosting Machine (GBM) is an advanced technique for predicting changes in cryptocurrency prices. It builds a series of decision trees, where each new tree aims to correct the mistakes of the previous ones. This step-by-step process helps create a strong model that can capture intricate patterns in the data. GBM works well with large datasets that have many variables, improving prediction accuracy by optimizing the model's performance through a method called gradient descent. This technique is particularly useful when there are non-linear interactions between sentiment and price changes, making it a powerful tool for complex prediction tasks.

4.4 Resampling

Different resampling techniques are used to explore the effectiveness of each method for predicting cryptocurrency price changes.

4.4.1 Simple Data Splitting

Simple data splitting involves dividing the dataset into two parts: a training set and a test set. The training set is used to build the model, while the test set is used to evaluate its performance. This method is straightforward and fast but may not be reliable if the data size is small or not representative

4.4.2 5 Fold Cross-Validation

Cross-validation is a technique where the dataset is divided into multiple subsets (folds). The model is trained on several folds and tested on the remaining fold. This process is repeated 5 times, with each fold serving as the test set once. The results are then averaged to provide a more reliable estimate of model performance. This technique helps in reducing bias and variance in the evaluation process.

4.4.3 Bootstrapping

Bootstrapping involves repeatedly sampling from the dataset with replacement to create multiple training sets. The model is trained on each set and evaluated on the original data or an out-of-bag sample. This technique allows for estimating the accuracy of the model and its variability. Bootstrapping is particularly useful for small datasets as it maximizes the use of available data.

By assessing these techniques, the most effective method for our research needs will be determined.

4.5 Model Evaluation

Two main performance metrics, RMSE as R-squared, are calculated to determine the effectiveness of the models in predicting cryptocurrency price changes. These metrics are necessary to understand how well our models predict changes in cryptocurrency prices.

R-squared measures how much of the variation in the price data is explained by our model. A higher R-squared value means the model's predictions are closer to the actual price changes, indicating a better fit.

RMSE, on the other hand, shows the average error in our model's predictions. It calculates the differences between predicted and actual values, with lower RMSE values indicating more accurate predictions. This metric helps us see how far off our predictions are on average.

Using both R-squared and RMSE gives a comprehensive view of the model's performance. This dual approach helps fine-tune the models, making them more effective in predicting the volatile cryptocurrency market.

4.6 Lag Features

As often market reactions to news or events are not immediate a lag analysis is going to be computed for the correlation, linear model, random forest.

Using various time lags to determine the optimal predictive lag period. The core of the analysis involves shifting the sentiment data backward and forward by different time periods to explore how previous and subsequent sentiments correlate with price changes. This shifting helps to test whether the market's reaction to sentiment is immediate or delayed.

- Sentiment scores are shifted backward to align with future price changes, hypothesizing that current sentiment affects future prices. Conversely, sentiment data is also shifted

forward to assess if previous price changes can predict future sentiment shifts, capturing the market's reactionary nature.

- For each lag configuration, the correlation between lagged sentiment and subsequent price changes is calculated. Models are then used to quantify this relationship, providing metrics such as the R-squared value and root mean squared error (RMSE), indicating the model's accuracy.

4.7 Predictions

First, predictions are generated from each model separately. Each model analyzes the data differently, focusing on the sentiment data and specific time lags that are most relevant for predicting price movements. This provides various perspectives on potential future outcomes.

After that, these individual predictions are combined into a single forecast. This gives a more comprehensive view of expected price changes. By merging the strengths of each model, the goal is to make the predictions more reliable and accurate. Each model has its own advantages and limitations so this approach allows to leverage the unique insights from each model, improving the robustness of the forecasts.

5 DATASET

The research uses a variety of datasets to explore the relationship between sentiment and cryptocurrency price.

First, an historical news dataset as well as a historical bitcoin price dataset from Kaggle [10] is used to get a past trends overview. This historical data contains Crypto news data from over a year (2021-10-12 / 2023-12-19) from three different sources : cryptonews.com, cryptopotato.com, cointelegraph.com. Three reliable crypto news media. This data initially came with pre-calculated sentiment scores. However, these were removed to use the FinBert model to generate sentiment scores. During the cleaning phase, the date formats were standardized to ensure compatibility with thr analytical function, and the news data was pared down to essential columns like 'date', 'headline', and 'description'. Then, the Bitcoin price data was retrieved from the CSV file, where the 'Date' column was cleansed of extraneous text and converted for uniformity. Only the 'coin name', 'timestamp', and 'price' columns were retained, thus standardizing this dataset for subsequent analysis.

Google Trends [11] data for an initial exploration were also used, which involved minimal cleaning such as converting the 'MONTH' column to a datetime format and setting it as the index. Another dataset came from Yahoo Finance [12], which included daily price data that we cleaned by removing missing values and standardizing the 'DATE' column. Then the price data was resampled to a monthly frequency to align with the Google search data, creating a dataset that combines closing prices, trading volume, and search trends.

In complement of that, further datasets for this research were used which are derived from a comprehensive collection of cryptocurrency-related news articles and financial data that are obtained through various APIs that provide both historical and real-time data concerning cryptocurrency market dynamics and news sentiment. The monthly range that will be mainly used is from 2024-04-18 to 2024-05-21 whereas the daily range that will be used is 2024-05-04 to 2024-05-06. However, other timestamp analysis are available in the *metrics.xlsx* in the github repository or via the different jupyter notebooks. The data, coming from APIs, came clean as it was directly processed during the fetching process. As this is part of another parallel project, the details will not be examined further here.

Those information form the backbone of the analysis, offering insights into the impact of market sentiment on cryptocurrency prices.

To refine the data further for analytical purposes, several data processing steps were implemented. Text fields were normalized by removing case sensitivity and trailing spaces and by filtering out duplicate entries. Dropping news articles that shared the same initial segment of the headline, ensuring the uniqueness of each record.

Cryptocurrency price data needed to be aligned with news data by matching each price entry to the nearest date in the news dataset. This precise alignment is needed for correlating market price movements with specific news events and for further sentiment analysis. Prices were normalized and aggregated between 0 and 1, enabling a direct comparison with sentiment scores derived from news content.

To calculate daily closing prices, we recorded the last price of each day, ensuring it reflects the day's final market sentiment. We then computed daily price changes as percentages to gauge day-to-day volatility and the influence of news sentiment on the market. We used the shift(-1) method to adjust the price change data backward by one day, aligning it with the corresponding day's sentiment data. This helps analyze how daily news sentiment could predict price movements.

The processed price data and sentiment scores were then merged based on date indices, forming a comprehensive dataset that includes key variables such as normalized prices, next-day price changes, and average daily sentiment.

The final dataset is reviewed for any missing values that might arise from non-overlapping

dates between the news and price data or the data shifting process. Such missing values are addressed to ensure the dataset's completeness, either by imputation of missing values or by excluding incomplete records, preparing the dataset for robust and reliable analysis.

6 IMPLEMENTATION

6.1 Sentiment Analysis

Having the datasets ready for the implementation, the *SentimentAnalyzer* class is used, with it's ProsusAI/FinBERT model.

The class initializes a tokenizer and model from the pretrained FinBERT model. If a GPU is available, the model is loaded onto it to speed up the computations.

The *predict_sentiment_batch* method processes the text data in batches. This is crucial for handling large datasets efficiently. Text data is then tokenized using the Bert Tokenizer, converted to tensors, and processed in batches. Each batch of text is then forwarded through the FinBERT model to obtain sentiment predictions. These predictions are represented as logits, which are converted into probabilities using the softmax function to provide a clearer interpretation of the sentiment values.

The *add_sentiments_to_df* method takes the calculated sentiment scores and integrates them into the original DataFrame by appending a new column. This allows the enhanced DataFrame to contain both the original data and the sentiment analysis results. This integration facilitates in-depth analytical reviews and straightforward visualization within the DataFrame structure.

The resulting DataFrame includes a new column of sentiment scores ranging from 0 to 1, where higher values indicate more positive sentiment. These scores provide a quantitative measure of the sentiment prevalent in each textual entry in the dataset, which can include financial reports, news articles, or social media blurbs [3].

6.2 Predictive Analysis

6.2.1 Sentiment and price

The Visualizations class offers graphical representations to get an intuitive look on how the data behaves. The packages *Matplotlib* and *Seaborn* are used for this data visualization due to their flexibility and ease of use. The function *plot_normalized_price_and_sentiment* is designed to plot both normalized prices and sentiment over time on a dual-axis chart. The dual-axis setup allows viewers to directly compare how changes in sentiment might correlate with shifts in cryptocurrency prices. This visual comparison is interesting to intuitively identifying potential relationships.

To make the sentiment trends clearer and less noisy, the sentiment data is smoothed using a rolling mean with a window defined by the *window_size* parameter. This smoothing helps in highlighting broader sentiment trends rather than reacting to abrupt fluctuations which may not be significant.

6.2.2 ARIMA forecast

The *forecast_prices_with_arima* function in the code is designed to utilize the Auto ARIMA model for forecasting future cryptocurrency prices based on historical data. the forecast_prices_with_arima function uses the Auto ARIMA model to predict future cryptocurrency prices using past data. The *pm.auto_arima* tool from the *pmdarima* package helps find the best settings by trying out different combinations of parameters. This function works with non-seasonal data, which is common in daily or monthly price information. It predicts prices for upcoming periods, useful for short-term trading. It also gives confidence intervals to show the possible range of future prices, helping in risk assessment. Additionally, the function creates a graph that displays these predictions alongside past prices to make comparison easy.

6.2.3 Feature Engineering – Lag

The analysis function begins by preparing the dataset based on specified time intervals and lag periods. Different lags of sentiment data are analyzed to determine their potential influence on future price changes, allowing us to identify the most predictive time frames.

The code defines a range of lags, both positive and negative, to shift the sentiment data backward and forward in time. For instance, lags are set up as [-7, -5, -2, -1, 0, 1, 2, 5, 7, 10], where negative numbers represent a backward shift (future prediction scenario), and positive numbers represent a forward shift (past analysis scenario).

These lags correspond to the unit of time defined by *from_date* and the time variable, which could be different depending on the granularity of the data being analyzed.

For each specified lag, the sentiment data is shifted using the *.shift()* method in pandas. This method is applied to the average sentiment column of the combined data DataFrame.

Negative lag shifts data backward in time to see if sentiment on a particular day influences price changes of the previous day, while positive lag shifts data forward to explore if sentiment on a given day reacts to the price change from the day before.

This shifted data (Lagged Sentiment) is then used in subsequent analyses to correlate and predict price changes.

6.2.4 Correlations

For each lag value, the function calculates the correlation between lagged sentiment and

subsequent price changes. This is achieved using pandas' *corr* function to compute the correlation coefficient, which helps determine the strength and direction of the relationship.

The iloc[0, 1] is used to access the correlation value of the off-diagonal element in the resulting correlation matrix, which represents the correlation between the two variables.

The calculated correlation coefficient for each lag is appended to a list correlations for further reporting.

6.2.5 Model Implementation and Execution

The dataset is then split using one of the resampling methods seen before, to evaluate the model's performance on unseen data.

The implementation code supports the different models seen before.

For the linear regression analysis, Ordinary Least Squares (OLS) function from the statsmodels Python library is utilized. This involves adding a constant to the features for OLS regression, fitting the model on the training set, and then making predictions on the test set. The sentiment scores, adjusted by the specified lag, serve as the independent variable. To accommodate the requirements of the OLS model from statsmodels, a constant is added to the array of independent variables using sm.add_constant(). This step is crucial as it allows the model to include an intercept in the regression equation.

For the random forest, a more complex approach where a RandomForestRegressor is trained. This model can capture non-linear dependencies better than linear models. The choice of n_estimators=1000 was made to ensure a robust and stable model by averaging the results of a large number of decision trees, which helps to reduce overfitting and improve predictive accuracy. The specific value of 1000 is often used and it strikes a balance between achieving these benefits and managing computational efficiency.

For the Gradient Boosting Machine the *GradientBoostingRegressor* from sklearn was used. This model handles non-linear relationships by building new. Configured with parameters such as n_estimators=100, learning_rate=0.1, and max_depth=3, to optimize performance while preventing overfitting.

Finally, for the support vector regression (SVR) the *SVR* function from *sklearn* was used. SVR is useful for datasets with complex, non-linear patterns that linear models can't capture. Setting the kernel to *rbf* to handle non-linearity, and configure the C=100 and gamma=0.1 parameters to adjust the regularization and the influence of a single training example, respectively, with *epsilon=.1* determining the margin of tolerance within which no penalty is given to errors.

All the model, were train on lagged sentiment scores, with price changes as the target.

The lagged sentiment scores serve as the input features. And the corresponding price changes are used as the target variable.

6.2.6 Performance Evaluation

After model training, key two metrics ,R-squared and RMSE, are calculated to assess the fit and predictive accuracy of the model. These metrics are needed for validating the model's effectiveness.

R-squared, Calculated using the *r2_score* function from *sklearn.metrics*, represents the proportion of the variance in the dependent variable that is predictable from the independent variable.

Root Mean Squared Error (RMSE), computed using the *root_mean_squared_error* function also from *sklearn.metrics*, is a measure of the accuracy of the model in predicting the dependent variable.

6.2.7 Future predictions

For the ARIMA model, percentage Change Calculation are computed from the last known price, alongside the upper and lower bounds of the confidence intervals. These statistics are aggregated to provide a summary of expected price movements and their potential range.

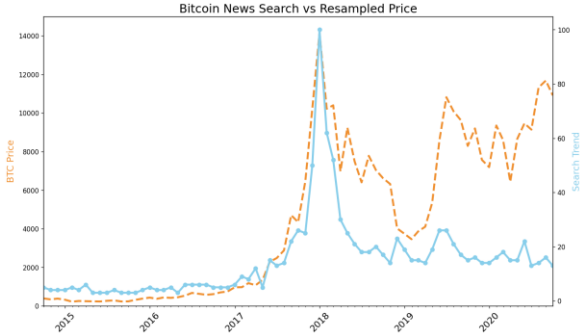
For the models, future price changes are predicted using the most up-to-date sentiment data, which is adjusted according to the optimal lags identified during the model training phase. Once the sentiment data is prepared, it is fed into the fitted models to forecast future price movements. This process uses the model's *.predict()* method, where the input features are the lag-adjusted sentiment scores.

7 RESULTS

The result of the sentiment analysis project illustrates several key insights into the relationship between news sentiment and cryptocurrency price movements.

7.1.1 *Google Trends*

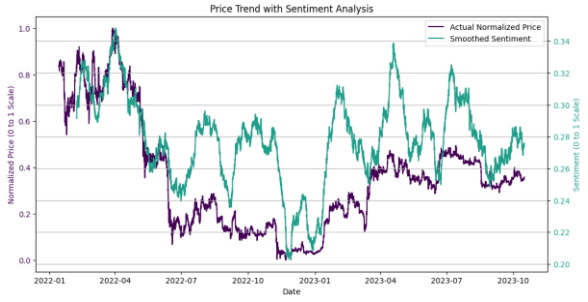
Firstly, this plot shows the relationship between Google search trends for Bitcoin (orange dashed line) and Bitcoin's price (solid blue line) from 2015 to 2020. Around late 2017, both search trends and Bitcoin prices peaked, with search interest rising slightly before the price. From 2018 to 2020, both metrics showed volatility, with Bitcoin's price fluctuating more than search trends. By late 2019 to 2020, both lines stabilized, indicating a steadier market and interest level.



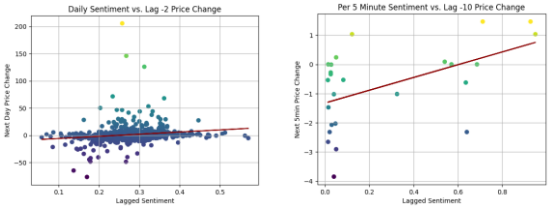
This suggests that higher search volumes might precede price changes, highlighting the influence of public interest on Bitcoin's market performance. With a correlation coefficient of 0.64, this preliminary analysis indicates that as sentiment increases, prices tend to rise as well but we see that this effect diminish through time. Maybe people have to much expectations or the currency becomes less volatile.

7.1.2 *Sentiment Using FinBERT*

In the analysis of historical datasets the early 2022 period shows fluctuations in cryptocurrency prices that are almost perfectly aligned with shifts in news sentiment. This intuitively makes us suppose that there is indeed a relationship where news events often coincided with movements in cryptocurrency prices, which suggest a potential correlation between news sentiment and market behaviour.

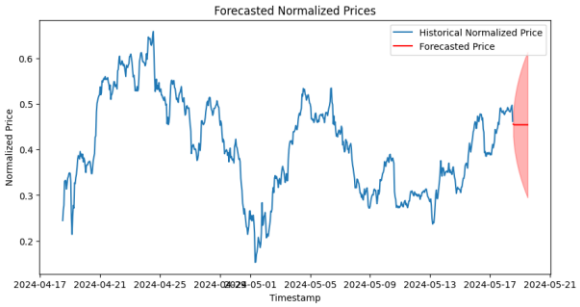


Throughout the analysis, when testing with different categories of coins and timestamp, a moderate correlation was observed, with different values from -0.5 to 0.5,. Thus, the overall correlation across various datasets remained weak and inconsistent, typically not surpassing 0.5. This suggests a lack of a strong, reliable predictive relationship between news sentiment and price changes. As seen here, correlation of 0.17 (left) with the 01-2022 to 10-2023 period and



a correlation of 0.5 (right) for the daily APIs Datasets

The ARIMA model frequently reverted to a naive forecasting approach, where the most recently observed price was projected forward. This tendency reflects the common characteristic of financial time series data, which often resemble a random walk. This provided however a good benchmark for further comparison through the RMSE metrics which was very low, often between 0.02 and 0.2.



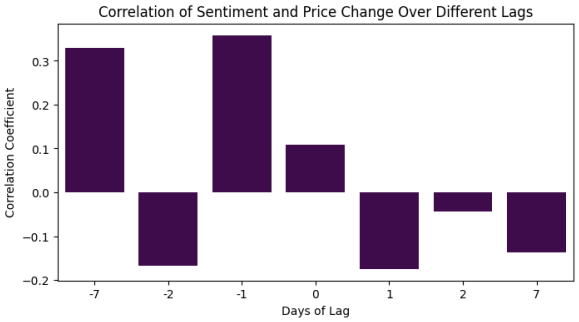
7.1.3 *Time Period*

It is also interesting to note that on a larger scale, when taking all the historical data from 01-2022 to 10-2023, the correlation where weaker and the metrics like R-square were almost always 0 between all resampling techniques. This indicates that the variability in the data may be too high for a simple correlation to capture the relationship between news sentiment and cryptocurrency price changes effectively. The results suggest that while individual events or short-term trends might momentarily align with sentiment indicators, the overall influence of news on market prices does not consistently hold over longer periods. Short-term analysis might therefore be preferable.

7.1.4 *Lag Features*

Additionally, the resampling and lag analysis techniques helped to uncover that the effect of sentiment on price changes is not immediate but rather delayed, with the strongest correlations observed at specific lags depending on the cryptocurrency and the nature of the news. Also, the various analysis showed that the most frequent lags that obtain relevant result where the minus lags, such as -7 or -2. Indeed, those lags provided the most positive correlation and

relevant metrics whereas the positive lags often show inverse correlation and bad metrics.



Although one might naturally think the market would react to news right away, it actually takes some time for the effects to show. This is probably because investors need a couple of days to fully understand the news, talk it over, and make smart choices about buying or selling, or, more likely, because most of the news are simply not that 'big'.

Furthermore, the data suggests an inverse relationship in most cases where higher sentiment correlates with lower prices immediately or shortly after, which is contrary to typical expectations. This might imply that high sentiment could be a reaction to peak prices which then correct downwards or this could imply overreaction in prices to positive sentiment, leading to *corrections*. This is a often reccurring pattern in the meme coin in crypto

7.1.5 Resampling Assessment

Historical 2022-01 to 2023-10 Analysis - BTC

Several evaluation of resampling methods was conducted to determine the most suitable approach for the analysis.

We are going to focus here on the timeframe 01-2022 to 10-2023 and BTC for this analysis. Note that the metrics can be compared in the *metrics.xlsx* file.

Initially, no resampling method was used. This provided, quite relatively high metrics especially for the random forest model and the GBM model.

For -2 Days Lag	corr	r.squared	RMSE
LR	0.17	0.03	4.02
RF	**-0.04**	**0.76**	**2.83**
GBM	**0.17**	**0.68**	**3.04**
SVM	0.17	0.01	4.04

This suggests that the random forest model or the GBM models captured complex data patterns effectively. However, this exceptional performance raises concerns about potential overfitting, as the model might be tailored too closely to the training data nuances. This is indicated by its ability to build deep, detailed trees, potentially fitting noise rather than just valid patterns. Further results for different timespan, coins category can be seen in the *no_splitting.ipynb.* Also, not too much 'bad' metrics like negative R-squared or tremendously high RMSE were observed.

Then, basic splitting was applied.

For -2 Days Lag	corr	r.squared	RMSE
LR	0.17	0.02	20.17
RF	**0.17**	**-0.11**	**21.51**
GBM	**0.17**	**-0.07**	**21.04**
SVM	0.17	0	20.38

Implementing simple data splitting, there was a dramatic decrease in R-squared values, dropping sometimes into negative territory. This highlights significant fit issues across all models.

This dramatic decrease suggested that the models might not only be failing to capture the underlying patterns in the split datasets but could be modelling the noise instead. Negative R-squared values indicate that the chosen models perform worse than a horizontal line representing the mean of the dependent variable, thereby not providing useful predictions. This problem highlights the inadequacy of the models when dealing with split data, which may not represent the full complexity or variability of the data as effectively as other resampling methods. Further results can be observed for other models, timestamp and coins categories in the splitting.ipynb

Afterwards, 5 Fold-Cross Validation was applied.

We observe a similar drop in all models metrics. Further results can be observed in the cv.ipynb

For -2 Days Lag	corr	r.squared	RMSE
LR	0.17	0.02	15.48
RF	**0.17**	**-1.32**	**20.4**
GBM	**0.17**	**-1.34**	**20.03**
SVM	0.17	0.02	15.55

Then, bootstrapping was applied, which increased the r.squared again but relatively lower than the no resampling method.

For -2 Days Lag	corr	r.squared	RMSE
LR	0.17	0.11	7.42
RF	**0.17**	**0.17**	**7.18**
GBM	**0.17**	**0.11**	**7.42**
SVM	0.17	0.17	7.18

This confirms the overfitting problem suspected previously. Further result can be observe in the *bootstrapping.ipynb*

API Monthly Data – Arpil 2024 – 50 Meme Coins

We are going to now focus on the timeframe month of April for the subsequent results analyiss.

Initially, similar to the BTC yearly data, without resampling, models like Random Forest and Gradient Boosting Machine (GBM) displayed relatively high performance, suggesting strong data pattern capture. Yet, this likely indicates overfitting, particularly for GBM, which achieved a near-perfect R-squared of 0.99

For -2 Days Lag	corr	r.squared	RMSE
LR	0.31	0.1	4.49
RF	0.31	0.83	2.97
GBM	**0.31**	**0.99**	**1.23**
SVM	0.31	0.02	4.58

Upon simple splitting, this problem is confirmed by significant performance drops when simple splitting is applied.

For -2 Days Lag	corr	r.squared	RMSE
LR	**0.31**	**-0.27**	**29.99**
RF	**0.31**	**-0.28**	**30.16**
GBM	**0.31**	**-0.45**	**32.14**
SVM	0.31	-0.11	28.08

5 Fold Cross Validation provided the same drop.

For -2 Days Lag	corr	r.squared	RMSE
LR	**0.31**	**-0.24**	**20.79**
RF	**0.31**	**-1.25**	**22.33**
GBM	**0.31**	**-2.74**	**25.53**
SVM	**0.31**	**-0.11**	**20.83**

Bootstrapping on the other hand appeared to stabilize performance somewhat, although inconsistencies remained, particularly for support vector machines (SVM), where R-squared remained low.

For -2 Days Lag	corr	r.squared	RMSE
LR	**0.31**	**0.46**	**15.91**
RF	**0.31**	**0.56**	**14.38**
GBM	**0.31**	**0.47**	**15.67**
SVM	**0.31**	**-0.01**	**21.69**

Therefore, while the initial results from the no-resampling method on the 50 meme coins dataset highlighted Random Forest's as a candidate for high accuracy, the more tempered results from bootstrapping provided a clearer, more realistic indication of its capacity to generate reliable and generalizable outcomes. This suggests that for small datasets or in applications where the data quality varies significantly, combining Random Forest with bootstrapping might offer a balanced approach, achieving both high performance and robustness in predictive modelling.

8 CONCLUSION

In conclusion, the results from this project demonstrate the potential of using automated sentiment analysis combined with historical price data to predict future price movements in the cryptocurrency market. However, it also highlights the complexity and challenges in accurately modelling and predicting such volatile and sentiment-driven markets.

For the linear regression low R-squared values and high RMSE across all models was observed which suggest that a linear approach may not be the best fit for predicting price changes based on sentiment. Similarly, SVM provided suboptimal metrics across models, further underscoring its limitations in this context. The lower RMSE values across all lags compared to those from the linear model suggest that the Random Forest model with bootstrapping resampling provides more accurate and reliable predictions. GBM was the second-best model, which can be explained by its ability to iteratively correct errors from previous trees, enhancing its accuracy in dynamic markets.

However, the stark contrast in performance upon implementing more complex resampling techniques such as cross-validation and basic splitting—where the metrics significantly worsened—suggests that the initial results might have been overly optimistic, hinting at overfitting.

Overfitting is a critical concern in machine learning, especially with algorithms capable of developing highly complex models like Random Forest and should be carefully addressed.

Bootstrapping revealed to be the best resampling method for this purposes. For small datasets, such as those used here, Random Forest paired with bootstrapping becomes particularly advantageous. Small datasets are more prone to overfitting because there's simply less data to learn from, making every single outlier or noise potentially more influential. By using bootstrapping, the model's exposure to different facets of the data is maximized without requiring more data than is available, effectively augmenting the dataset's size and diversity in a virtual sense.

The more recent and short time period (April 2024/50 meme coins), found through the API, provided better results, potentially comprising more uniformly relevant or 'better' news than the news obtained in the Kaggle Dataset. Indeed the RMSE was closer to the ARIMA model which further indicates it’s reliability.

Furthermore, the lag analysis outlined a delay in the market response to news. It's therefore important not only to look at what the news says but also to consider when the market starts to respond to it.

The findings suggest that while sentiment is a significant factor, variability exist and sometimes sentiment might not be the preponderant factor. This increase even more the notion of transparency, that we discussed previously in the introduction part, in sentiment analysis approach to price prediction.

The limitations of this project are that, it used free tier API data, which might be less dense than paid news sources, and way simpler. This resulted in small datasets to use, which might have skewed the results. Further analysis like fighting overfitting better, using more complex ensemble techniques, and integrating richer datasets could enhance future studies. Additionally, the computational resources available were insufficient to perform more robust validation techniques such as 10-fold cross-validation or extensive hyperparameter tuning. This research did however provide some insights for more refined models and suggests pathways for addressing the inherent challenges of predictive analytics in the cryptocurrency space.

9 REFERENCES

[1] L. Chappex, "Interview with Richard Peterson, CEO of MarketPsych," Swissquote, [Online]. Available: https://en.swissquote.lu/international-investing/investing-ideas/interview-richard-peterson-ceo-marketpsych. [Accessed: May 21, 2024].

[2] L. Chappex, "Market mood dissected by AI," Swissquote, [Online]. Available: https://www.swissquote.com/en-ch/market-mood-dissected-ai. [Accessed: May 21, 2024].

[3] B. Fazlija and P. Harder, “Using financial news sentiment for stock price direction prediction,” *Mathematics*, vol. 10, no. 13, p. 2156, 2022. [Online]. Available: https://doi.org/10.3390/math10132156

[4] T. Adams, A. Ajello, D. Silva, and F. Vazquez-Grande, “More than words: Twitter chatter and financial market sentiment,” *Finance and Economics Discussion Series*, vol. 2023-034, Board of Governors of the Federal Reserve System, Washington, 2023. [Online]. Available: https://doi.org/10.17016/FEDS.2023.034

[5] W. Souma, I. Vodenska, and H. Aoyama, “Enhanced news sentiment analysis using deep learning methods,” *Journal of Computational Social Science*, vol. 2, pp. 33-46, 2019.

[6] M. P. Cristescu, D. A. Mara, R. A. Nerișanu, L. C. Culda, and I. Maniu, “Analyzing the impact of financial news sentiments on stock prices—a wavelet correlation,” *Mathematics*, vol. 11, no. 23, p. 4830, 2023. [Online]. Available: https://doi.org/10.3390/math11234830

[7] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," IEEE Access, vol. 8, pp. 131662-131682, 2020.

[8] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," Expert Systems with Applications, vol. 41, no. 7653-7670, 2014.

[9] D. T. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” *arXiv preprint arXiv:1908.10063*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1908.1006

[10] Search | Kaggle

[11] Google Trends

[12] https://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD