TITLE: Instructions for the Authors of Papers

Urs A. Hurni

University of Lausanne

Lausanne, Switzerland

urs.hurni@unil.ch

# Abstract

# 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 realm of modern finance, the role of sentiment analysis becomes increasingly crucial. By harnessing sophisticated algorithms and machine learning techniques, analysts and investors are now able to decipher 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

# 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?*

## 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.

## Objective

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.

## Literature Review

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]

## Scope

For the further analysis we are going to use a set of a pre-loaded dataset, only for testing purposes. That looks like that : For the 24h period : 2024-05-18\_24h\_news\_with\_sentiment.csv and 2024-05-18\_24h\_50meme\_history.csv and for the 30 day period : 2024-04-05\_30d\_news\_with\_sentiment.csv and 2024-04-05\_30d\_50meme\_history.csv.

These datasets are representative of the type of data typically accessible through APIs. However, 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.

# Methodology

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

## Data Collection

To thoroughly encompass the diverse and nuanced landscape of cryptocurrency news, it is crucial to gather information from a broad and varied dataset. Therefore, five specialized crypto news APIs have been selected as the primary sources.

This includes NewsAPI, CryptoPanicAPI, CryptoDataFetcher, CryptoNewsAPI, and SeekingAlphaNewsAPI.

* NewsAPI: Provides broad news coverage including general updates in the cryptocurrency sector.
* SeekingAlphaNewsAPI: Known for its detailed financial analysis, this API provides insights focused on cryptocurrency market movements and significant financial trends that could impact the cryptocurrency landscape.
* CryptoPanicAPI, CryptoDataFetcher, and CryptoNewsAPI: These APIs have been chosen for their availability and to enrich the dataset with targeted cryptocurrency news and timely updates, ensuring a comprehensive collection of information directly relevant to the crypto markets..

To enhance this global perspective, adding an economic perspective is interesting. Therefore, USEconomyAPI will be used for this macro economic perspective

* USEconomyAPI: Offers news articles focused on economic aspects that might impact or reflect the state of the broader economic markets.

Furthermore, one crypto price API will be used for gathering real time price data.

* Coinkranking API, provides real-time cryptocurrency pricing as well as historical pricing.

Different classes for each APIs are used to fetch and retrieve news articles related to cryptocurrencies and economic context for adaptability and ease to use. They provide a rich and up-to-date dataset of textual content which is then analyzed for sentiment.

## Data Processing

Each dataset fetched from API will need to be transformed to match the same format and to allow further usage. Indeed, every API outputs it’s data in different format. The DataHandler class (hd) plays a central role in managing the various tasks associated with data retrieval, processing, and storage. This section provides an overview of the data processing activities, with further details to be covered in subsequent sections.

1. All dates and header needs to be in the same formats. Thus, standardization is needed. Crucial to maintain consistency.
2. The datasets obtained from the APIs are merged into a single DataFrame. This consolidation is necessary to combine varied data points into a unified format for comprehensive analysis.
3. The merged data is then cleaned, this involves processing duplicates by comparing text fields across news articles
4. Cryptocurrency price data is transformed to align with news data as the merge occurs via the ‘date’.
5. Price data is scaled between 0 and 1 to match the scale of sentiment scores derived from news articles, facilitating comparative analysis.
6. Using resampling techniques, daily closing prices are computed by taking the last recorded price for each day, which is at 12PM.
7. Price changes are calculated as daily percentages to analyze the relationship between market movements and news sentiment.
8. To study the effect of news sentiment on subsequent price movements, the daily price change data is shifted backward by one day (shift(-1)), aligning it with the corresponding day’s sentiment data.
9. The processed price data and sentiment scores are merged based on date indices. This merged dataset includes key variables such as normalized prices, the following time period price change, and average time period sentiment. This combination is critical for examining the interplay between market sentiment derived from news and actual market performance.
10. The final dataset is reviewed for any missing values resulting from non-overlapping dates or the data shifting process. Rows containing NaN values are carefully handled to ensure the dataset's completeness and readiness for robust analysis.

## Model Application and Description

### 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.

### Correlation

This 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.

### Linear Regression

Linear regression in this context utilizes the Ordinary Least Squares (OLS) method from the statsmodels library to forecast cryptocurrency price changes based on sentiment scores. 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.

### 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.

## Splitting

Linear regression and random forest models are split into a training set, comprising 75% of the data, and a test set, containing the remaining 25%. This division is designed to ensure a more realistic simulation of model performance in real-world scenarios. By training the models on the larger portion of the data, they can learn and adapt to the underlying patterns and relationships between sentiment scores and cryptocurrency price changes. The smaller test set then serves as a new, unseen dataset for evaluating how well these models can generalize their predictions to new data. This approach helps to mitigate overfitting and provides a clear measure of model efficacy and predictive accuracy outside of the training sample.

## Model Evaluation

Using the previously described data split, various performance metrics such as R-squared, mean squared error (MSE), and accuracy are calculated to determine the effectiveness of the models in predicting cryptocurrency price changes. These metrics are crucial for assessing the precision and reliability of the predictions made by the models.

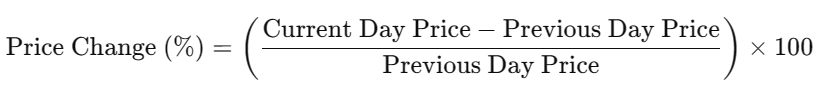
To provide a thorough evaluation of model performance, two key metrics are computed: R-squared and Root Mean Squared Error (RMSE). R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. This metric is especially useful in comparing the fit of different predictive models, as it provides a sense of how well the unseen data points are replicated by the model. A higher R-squared value indicates a better fit to the data.

Root Mean Squared Error (RMSE), on the other hand, measures the average magnitude of the errors between predicted and observed values, providing a clear indication of prediction accuracy. Unlike R-squared, RMSE offers more direct insight into the actual differences in predicted vs. actual values, with a lower RMSE reflecting higher accuracy.

Together, R-squared and RMSE provide a comprehensive view of model performance, highlighting both the accuracy and efficiency of the models in terms of their predictive capabilities. By analyzing these metrics, we can fine-tune the models to improve their predictive accuracy, adjust the complexity of the model if necessary, and better understand the dynamics of the factors influencing cryptocurrency prices.

# A description of the data set (where it comes from, how it was cleaned, etc...)

The dataset for this research is derived from a comprehensive collection of cryptocurrency-related news articles and financial data, spanning multiple sources. The data is obtained through various APIs that provide both historical and real-time data concerning cryptocurrency market dynamics and news sentiment. This information forms the backbone of the analysis, offering insights into the impact of market sentiment on cryptocurrency prices.



# Implementation

## Fetching and Processing Data

Multiple API classes (NewsAPI, CryptoPanicAPI, CryptoDataFetcher, etc.) fetch cryptocurrency prices and news from various sources. For instance, CryptoDataFetcher retrieves historical price data using the Coinranking API, which is pivotal for analyzing the impact of news sentiment on price movements.

## Sentiment Analysis

The sentiment analysis is performed using a pretrained BERT model (Bidirectional Encoder Representations from Transformers) specifically tailored for financial sentiment analysis.

Sentiment Analysis

The SentimentAnalyzer uses the BERT model (via the transformers library) to compute sentiment scores from news headlines, which are later correlated with price changes to explore potential predictive power.

Integration and Execution

The **SentimentAnalyzer** class uses the ProsusAI/FinBERT model. FinBERT is a language model based on BERT, further pre-trained on a financial corpus to better handle the unique vocabulary and expressions used in financial texts [9]. This model is fine-tuned for financial sentiment analysis, making it particularly suitable for analyzing sentiment in cryptocurrency-related news articles.

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 tokenized, converted to tensors, and processed in batches. If a GPU is available, the tensors are moved to the GPU. The model predicts sentiment scores, which are then converted into probabilities using a softmax function. These probabilities represent the sentiment scores for each text entry.

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].

## Visualisation

Visualizations: Utilize the Visualizations class to create graphs and charts that illustrate the relationship between sentiment and price changes, model performance, and other relevant insights.

Reporting: Generate comprehensive reports that summarize findings, model performance, and potential strategies for leveraging sentiment analysis in trading.

By adhering to this methodology, the research aims to provide a thorough understanding of the dynamics between market sentiment and cryptocurrency prices, offering insights that could benefit investors and traders in making more informed decisions.

Visualization and Analysis

The relationship between cryptocurrency prices and sentiment data is visualized using the Visualizations class. This class processes and plots the data:

Normalization of Prices:

Prices of various cryptocurrencies are normalized to a scale from 0 to 1. This allows for fair comparisons across different cryptocurrencies.

Data Aggregation:

Sentiment and price data are grouped into regular intervals (e.g., hourly or daily). This ensures that each sentiment score directly corresponds to a price point.

Plotting:

The plot\_normalized\_price\_and\_sentiment function creates dual-axis charts that display both price and sentiment over time. These visualizations help identify whether changes in sentiment align with shifts in price trends.

## Predictive Analysis

Predictive models are applied to forecast future price changes based on sentiment data:

## Feature Engineering - Lag

To explore the potential time-delayed effect of sentiment on prices, lag features are created where sentiment scores are shifted by various time intervals (e.g., 1 day, 3 days, 7 days). This helps in identifying the optimal time lag that could predict price movements.

Lag Analysis

The code performs a comprehensive analysis of the relationship between market sentiment and cryptocurrency price changes using various time lags:

Shifting Sentiment Data:

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.

Correlation and Regression Analysis:

For each lag configuration, the correlation between lagged sentiment and subsequent price changes is calculated. Linear regression models are then used to quantify this relationship, providing metrics such as R-squared and RMSE.

Visualization and Output for Web:

Relationships and correlations are plotted, optionally generating web-optimized visualizations (SVG format).

Predictive Analysis:

The code uses derived models to forecast future price changes based on recent sentiment data, shifted by the analyzed lags. These predictions are aggregated to provide a comprehensive outlook on expected price movements.

# Results

The result of the sentiment analysis project illustrates several key insights into the relationship between news sentiment and cryptocurrency price movements. Upon implementing the various data fetching, processing, and visualization techniques, we observed noticeable trends and patterns in the data.

Firstly, the visualizations generated from the Visualizations class clearly highlighted the fluctuations in cryptocurrency prices in response to shifts in news sentiment. These plots showed that significant news events often coincided with sharp movements in cryptocurrency prices, suggesting a potential correlation between news sentiment and market behavior.

Through statistical analysis facilitated by statsmodels, we quantitatively assessed the correlation and found a moderate relationship in certain instances, particularly during high-impact news days. For example, during the release of major regulatory news or breakthrough technological advancements in blockchain, there was a noticeable impact on prices, which often moved in tandem with sentiment shifts.

Moreover, the sentiment analysis performed by the SentimentAnalyzer class, using advanced NLP techniques and the BERT model, provided a robust framework for understanding the emotional tone of news articles. The sentiment scores obtained were then correlated with price changes, revealing that particularly negative or positive news had a more pronounced effect on price volatility.

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.

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 modeling and predicting such volatile and sentiment-driven markets. The findings suggest that while sentiment is a significant factor, it must be integrated with other market indicators for more reliable predictions.

# 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.1908.10063