UFCFWQ-45-M
Interdisciplinary Group Project

Bitcoin Price Fluctuation Prediction Using Sentiment Scores of Newspaper Headlines

Nasir Dalal
Idil Ersudas
Umut Eyidogan
Josie Graham
Elizabeth Ofosu

Introduction

Bitcoin, a decentralised digital currency, was designed to prevent financial crises similar to the 2008 credit crunch through its self-governing nature and limited supply. Addressing issues related to online transactions and e-commerce growth, Bitcoin's blockchain technology ensures that each block must be confirmed before the next transaction, creating a traceable record of coin movements. The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work (Nakamoto, 2008). However, the lack of a centralised authority managing funds poses a significant risk; users losing access details can result in the permanent loss of their bitcoins.

A key feature of Bitcoin is its halving event, occurring approximately every four years, which reduces the reward for mining new blocks. This reduction in supply often leads to significant market rallies and media hype, creating a crypto bubble environment before eventually crashing into what is known as the crypto winter (Badev and Chen, 2014).

Bitcoin has gained negative attention due to its use in enabling criminal activities such as money laundering and the purchase of illegal items and services (Foley, Karlsen and Putniņš, 2019). Additionally, media hype often fuels FOMO (fear of missing out) and fearmongering through sensationalist headlines, exacerbating public concerns (Yermack, 2015). As a non-tangible currency with an unknown creator, Bitcoin remains a challenging concept, raising fears of scams. Created by the anonymous programmer(s) known as Satoshi Nakamoto, Bitcoin emerged on January 3, 2009, with the mining of the genesis block (Nakamoto, 2008). Despite its volatility and scepticism, Bitcoin has continued to gain acceptance and legitimacy. Notably, in 2021, Bitcoin was adopted as legal tender in El Salvador (BBC News, 2021), and in January 2024, the US approved 11 Bitcoin spot ETFs (Exchange-Traded Funds), solidifying crypto as a new asset class for investment (BBC News, 2024).

This study aims to predict whether to hold, sell, or buy Bitcoin based on newspaper headlines through sentiment analysis and machine learning. It utilises 15,592 headlines collected from February 25, 2018, to June 7, 2024. Sentiment analysis was conducted using the TextBlob and VADER libraries, with a comprehensive visual analysis highlighting their differences and similarities. The analysis seeks to explain these disparities and evaluate the most suitable library for our model.

Various machine learning algorithms were trialled during model development, including Support Vector Machine (SVM), Logistic Regression, and Random Forest, implemented via the scikit-learn library. The data followed an 80:20 training-to-testing split and was normalised using feature scaling. Technical indicators were created using the TA-Lib library to provide additional features for the machine learning models.

The models were evaluated using the test set and measured for accuracy, precision, recall, and F-Score. Additionally, a simulation was set up to compare the model's performance in a real trading situation to random choice, aiming to validate its practical applicability.

Literature Review

Bitcoin price prediction is a trending subject, and studies are consistently carried out to forecast the volatile nature of Bitcoin prices. This literature review examines recent studies exploring various methodologies for predicting Bitcoin prices, focusing on data preprocessing, sentiment analysis and machine learning techniques to be adapted throughout the study.

i. Preprocessing Techniques

A critical aspect of sentiment analysis and price prediction models is the preprocessing and data cleaning stage, as highlighted by several studies. Hossain et al. (2021) emphasised a comprehensive approach in their research on correlating sentiment between tweets and

newspaper headlines. Their method involved lowercasing text, removing punctuation and stop words, standardising abbreviations, and addressing spelling differences - particularly crucial when dealing with American and British English sources. They also opted for sentence-level tokenisation to preserve contextual integrity.

Moreover, Colianni et al. (2015) who applied Naive Bayes for cryptocurrency trading based on Twitter sentiment analysis found that including stopwords in their analysis improved the model's performance, contrary to common practice in natural language processing. They achieved F1 score of 0.53 and 0.56 without stopwords and with stopwords respectively. These findings emphasise that what works as a common rule may only sometimes be ideal for specific applications, especially in specialised fields like cryptocurrency trading, where unique language patterns may emerge.

Hence, Khyani et al. (2021) further contributed to this discussion by comparing lemmatisation and stemming techniques. They advocated for lemmatisation in projects dealing with concise text like headlines, as it preserves meaningful words more effectively than stemming. They mentioned the NLTK WordNetLemmatizer class and its Morphy function as practical tools for finding word roots. These preprocessing techniques are especially relevant for headline analysis, where each word can significantly impact sentiment scores.

The importance of such thorough data cleaning is also evident in earlier studies like Sattarov et al. (2020) and Nguyen Phuong et al. (2024). However, they provided less detailed preprocessing steps in their Twitter-based sentiment analyses for Bitcoin price prediction. These studies underscore that proper data cleaning and standardisation form the foundation for accurate model training and reliable predictions in cryptocurrency price forecasting and sentiment analysis.

ii. Sentiment Analysis

Two predominant types of sentiment analysis commonly used to predict Bitcoin prices are social media and news-based analysis, with social media being the most prevalent. Pagolu et al. (2016) found strong correlations between Twitter sentiment and stock price movements. Nguyen Phuong et al. (2024) extensively studied the impact of Twitter posts on Bitcoin price trends using sentiment analysis machine learning models. They compared Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM) models. The GRU model outperformed others, achieving a maximum accuracy of 90.3% at a 16-hour lag. In an earlier study, Sattarov et al. (2020) reported lower accuracy findings of 62.48% when predicting based on BTC-related tweet sentiment and historical BTC price.

Moreover, Zi Ye et al. (2022) combined lexicon-based sentiment analysis with LSTM models to predict Bitcoin prices from September 2017 to January 2021. Their results demonstrated that incorporating sentiment scores from social media comments reduced overall errors in price predictions. Roni et al. (2023) conducted sentiment analysis for Bitcoin, Ethereum, and Binance cryptocurrencies using Twitter data. They employed text mining and K-means clustering algorithms, highlighting the effectiveness of text preprocessing steps and sentiment analysis in predicting market movements.

Furthermore, studies that look at News-based sentiments studies include Gurrib & Kamalovs (2021), who aimed to predict next-day Bitcoin prices using a linear discriminant analysis (LDA)--based classifier that incorporated current BTC price information and news headlines. Chee Kean Chin and Nazlia Omar (2020) combined lexicon-based sentiment analysis of news articles with LSTM models to predict Bitcoin prices from September 2017 to August 2019. Their results showed that including sentiment scores from news articles reduced overall prediction errors.

Another study, Vo, Nguyen, and Ock (2019) examined a methodology for predicting Ethereum prices by combining news sentiment and historical price data, used a two-stage approach as sentiment analysis and price prediction in their study. In the sentiment analysis

stage, they employed natural language processing tools, including dependency parsing, co-reference resolution, and named entity recognition, to process news data and generate semantic and syntactic vectors. The vectors were combined to form an input vector that captures semantic and sentiment similarities among words.

iii. Machine Learning Models for Price Prediction

Studies have employed different machine learning algorithms, and no perfect model to predict Bitcoin price has yet to be developed. This is because cryptocurrency price forecasting is difficult due to price volatility and dynamism (Rathore et al., 2022). Consequently, the rapid changes in market patterns, which diminish prediction accuracy and the inability of models to precisely mirror past data trends, remain significant challenges.

Chen (2023) employed random forest regression and LSTM neural networks utilising 47 explanatory variables across eight categories from 2015 to 2022. This research reported that random forest regression slightly outperforms LSTM in prediction accuracy when compared between the two, though not statistically significantly. The best variable for the prediction of Bitcoin price was Bitcoin's previous day's OHLC prices, with other influential factors varying between the two studied periods (2015-2018 and 2018-2022). The study reveals that models using only the previous day's data perform best, supporting the efficient market hypothesis. While random forest regression shows promise for Bitcoin price prediction, it has limitations in forecasting prices beyond historical highs.

Slightly higher predictions of random forest were reported in other studies that were also conducting sentiment analyses. Pant et al. (2018) analysed Twitter sentiment to predict Bitcoin prices using a recurrent neural network (RNN) approach. For price prediction, the authors utilised a Recurrent Neural Network (RNN) model, explicitly mentioning variations like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The RNN model incorporated historical price data and sentiment scores to predict future Bitcoin prices. Their approach achieved a prediction accuracy of 77.62%.

Furthermore, Vo, Nguyen, and Ock (2019) used a Long Short-Term Memory (LSTM) network alongside historical price data to forecast future prices. The model was trained using 80% of the historical price data, with the remaining 20% reserved for testing. Performance metrics such as mean absolute percentage error (MAPE) and mean absolute normalised error (MANE) demonstrated the model's effectiveness in predicting price movements. This study underscored integrating sentiment analysis with time series data to enhance cryptocurrency price prediction accuracy.

In their research, Cocco et al. (2021) reported that Bayesian neural networks performed best for sentiment analysis and tweet volumes.

Qichuan Huang (2015) focused on predicting Bitcoin price movements using the Fear and Greed Index to measure market sentiment. The study employed machine learning algorithms, including linear regression, random forest, and XGBoost, with Grid Search XGBoost analysis determining the best model.

This literature review highlights the diverse approaches to Bitcoin price prediction, emphasising sentiment analysis of social media and news data. The literature review shows that machine learning models, particularly neural networks like LSTM and GRU, have shown promising results in combining sentiment data with historical price information. Also, effective preprocessing techniques cannot be overstated, as they form the foundation for accurate model training and reliable predictions.

Methods

All analyses were conducted using the Python programming language, leveraging the powerful data manipulation capabilities of the Pandas and NumPy libraries. These libraries

facilitated efficient data cleaning, transformation, and analysis, ensuring the robustness and reliability of the study's findings.

For the purposes of this study, sentiment scores of the newspaper headlines related to Bitcoin have been combined with Bitcoin price data with the aim of predicting future price fluctuations.

1. Bitcoin Price Data

Bitcoin prices were gathered using coincodex.com website. Dataset was stored as a CSV file., so the price data was collected with this accordance.

2. Newspaper Headlines

a. Data Collection

Two distinct data sources have been utilized in the data collection process: Nexus and Kaggle. Nexus is a comprehensive platform for accessing a wide range of news articles and publications (LexisNexis, n.d.), while Kaggle is a well-known data science platform that hosts various datasets and competitions (Kaggle, n.d.). For this project, Nexus was used to gather a large dataset of newspaper headlines related to Bitcoin by manually defining the required columns of data to ensure the dataset remained clean. Kaggle was used to download the pre-processed "Crypto News Headlines & Market Prices by Date" dataset uploaded by Aaron Bastian (Bastian, n.d.). "Bitcoin" was the only keyword used while searching for related headlines in these sources. The headlines published after 25.02.2018 and before 07.06.2024 were included in the analyses. A total of 15,592 headlines related to "Bitcoin" were discovered in the data collection process. When grouped by days, a total of 1,982 days were included in the analysis.

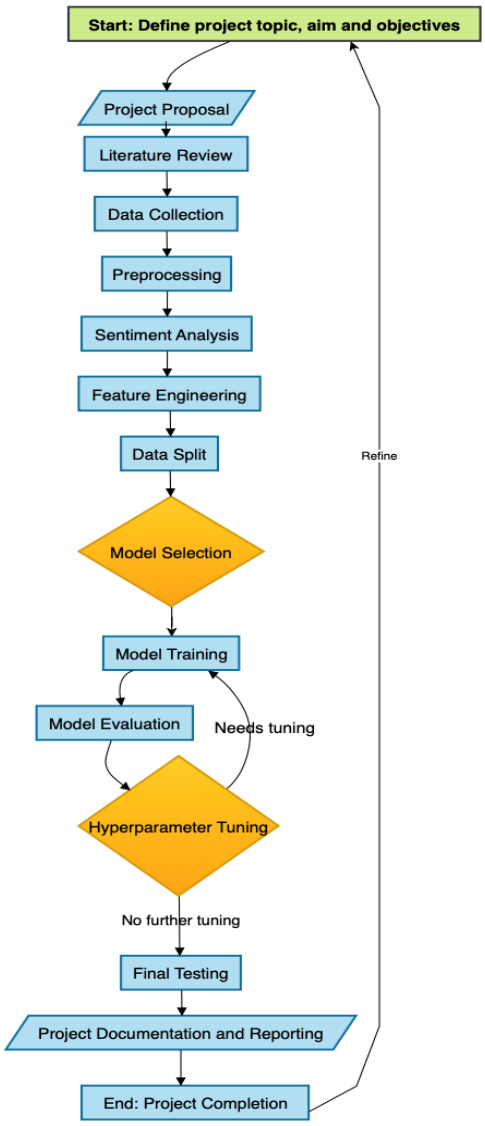

Figure 1: Proposed workflow.

b. Preprocessing

The data collected from Nexus was relatively clean, as the columns were specifically selected to include only the required information during the download process. However, the dataset from Kaggle included some additional columns related to Bitcoin prices that were not available in the Nexus data. Although Bitcoin price data is relevant for the purposes of this study, it was decided that, for the sake of consistency and reliability, Bitcoin prices should be gathered from a single source. Therefore, the additional columns from the Kaggle dataset were cleansed to maintain uniformity in the data and relevant action was taken to ensure the datasets can be combined. Additionally, the data was labelled based on the daily change in Bitcoin's closing price using the following code. Labels were assigned as follows: 0 = Significant Decrease, 1 = Decrease, 2 = Neutral, 3 = Increase, 4 = Significant Increase.

Initially, the dataset was cleansed of any headlines that were not written in English. Missing data were filled with NA values, and rows with null values for headlines were removed. To ensure uniformity, all headlines were converted to lowercase. The dataset was then examined for

duplicate entries, and any duplicates found were removed. Additionally, punctuation was stripped from the headlines to minimize noise in the dataset (Arjmand, 2024)

Stopwords, defined as commonly used words in a language that carry minimal meaning in the context of natural language processing (NLP) analyses (Blanchard, 2007), were removed using the common stopwords defined in the NLTK (Natural Language Toolkit) library (Sarica, 2021). To further reduce noise, the frequency of words in the headlines was analysed, and any additional dataset-specific stopwords deemed irrelevant to the analysis were also removed (Lison, Mowatt & Tjong Kim Sang, 2019;).

After successfully removing stop words from the headlines, the remaining data was checked for relevance to the topic using the BeautifulSoup library. BeautifulSoup, a widely used Python library, facilitates the extraction of data from web pages (Franc, n.d., Sarica 2021). For this project, BeautifulSoup was employed to scrape data from an online crypto dictionary (CryptoNest, n.d.). Each word in the headlines was compared against the crypto-relevant word list from the online dictionary to ensure that all retained headlines were pertinent to the topic.

Once it was confirmed that all headlines were relevant to the study, the remaining headlines were lemmatized (i.e., reverted to their root forms) using the TextBlob library (Sarica, 2021). This lemmatization process was performed at the end of the data cleaning phase, as it requires significant computational resources. Ensuring that all headlines were in their optimal form before lemmatization helped in maintaining computational efficiency (Arjmand, 2024).

c. Sentiment Score Calculation

The sentiment score calculation was performed using two Python libraries: TextBlob and VADER (Valence Aware Dictionary for sEentiment Reasoning). Both libraries were employed to conduct parallel analyses, ensuring robustness and reliability in the sentiment analysis process.

Initially, a custom lexicon specific to cryptocurrency was developed (CoinMarketCap, 2024; CoinGecko, 2024; Creatonics, 2023). This lexicon included terms such as "bullish," "bearish," "hodl," and other crypto-related jargon, each associated with a sentiment score. The purpose of this lexicon was to enhance the accuracy of the sentiment analysis by accounting for domain-specific terminology that might not be effectively captured by generic sentiment analysis tools.

Following the creation of the lexicon, sentiment analysis was performed on the collected newspaper headlines. TextBlob, a Python library for processing textual data, provides a simple API for executing common natural language processing (NLP) tasks (Aljedaani et al., 2022; Sarkar, 2019). This library was utilized to calculate the sentiment polarity of each headline. TextBlob's built-in sentiment analysis capabilities were supplemented by adjusting the scores based on the presence of terms from the custom crypto lexicon. This adjustment ensured that headlines containing crypto-specific terms were accurately scored according to the sentiment associated with those terms.

In parallel, the VADER library was used to perform sentiment analysis on the same set of headlines. VADER is particularly effective for analysing social media and other informal text, making it a suitable choice for evaluating the sentiment of news headlines (Hutto and Gilbert, 2014). This sentiment analysis tool provided a compound score for each headline, reflecting the overall sentiment as positive, negative, or neutral.

To ensure consistency and facilitate comparison, the sentiment scores from both TextBlob and VADER were labelled as positive, negative, or neutral. This labelling process involved defining thresholds for sentiment polarity scores, allowing the assignment of sentiment labels that could be aggregated and analysed.

The final step involved aggregating the sentiment scores by publication date. This aggregation provided a daily sentiment trend, highlighting the number of positive, negative,

and neutral headlines for each day within the analysis period. The aggregated sentiment data was then used to calculate the percentages of positive, negative, and neutral sentiments, offering a clear view of the overall sentiment trends over time.

By employing both TextBlob and VADER, the sentiment analysis leveraged the strengths of each library, ensuring a comprehensive and accurate assessment of the sentiment conveyed in the Bitcoin-related newspaper headlines. This dual approach provided a robust foundation for subsequent analyses and the development of predictive models based on sentiment and Bitcoin price data.

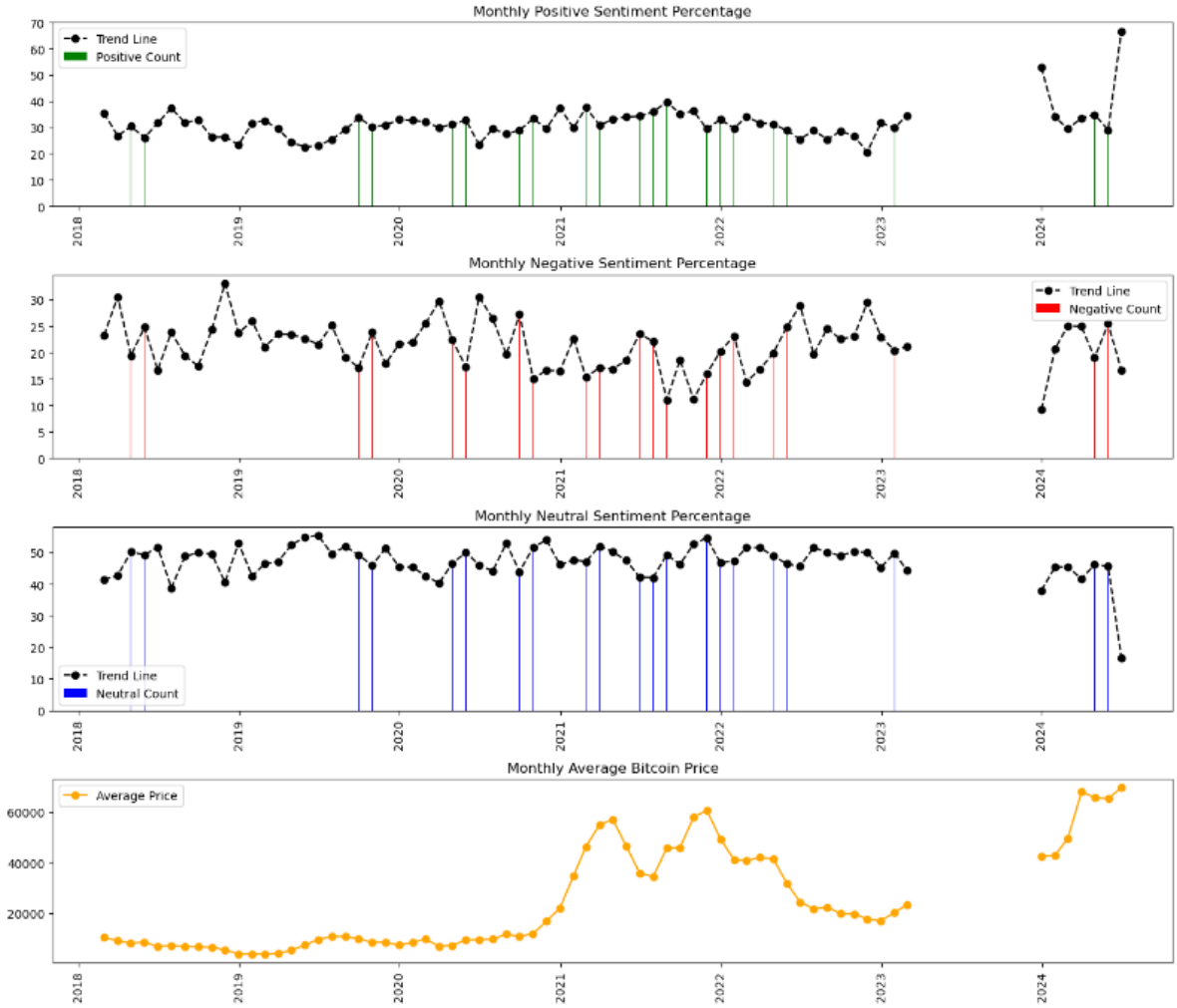

Figure 2: Monthly Sentiment Analysis and Bitcoin Price Trends.

Explanatory Data Analysis (EDA)

Data visualization was an integral part of this study to effectively communicate the results of the sentiment analysis and its correlation with Bitcoin price data. Various Python libraries such as Matplotlib and Seaborn were employed to create visual representations of the data. The visualizations included line plots, bar charts, and histograms to illustrate the trends and distributions of sentiment scores over time.

i. ***Sentiment Trends and Bitcoin Price Over Time****:*

Figure 2 provides a detailed visual presentation of monthly sentiment analyses related to Bitcoin and the average monthly price changes of Bitcoin from 2018 to 2024. Each graph displays the positive, negative, and neutral sentiments, alongside the average monthly price of Bitcoin as a time series, with trend lines added for each sentiment category. These trend lines

assist in identifying long-term trends, while the graphs themselves offer insights into monthly fluctuations.

In 2018, the overall sentiment regarding Bitcoin was mostly neutral. However, in 2019, there was a noticeable increase in positive sentiment, which continued until early 2021. During this period, Bitcoin's price generally rose, suggesting that positive news had a beneficial impact on the market. In 2021, an interesting pattern emerged where Bitcoin's price sharply increased throughout the year, but the sentiment analysis did not show a clear trend. Mid-2021 saw a rise in negative sentiment, which coincided with a price drop, indicating the market's sensitivity to negative news. By 2022, the decline in positive news and fluctuations in negative news mirrored the general downward trend in Bitcoin's price. This highlights how market sentiment and external news can significantly influence market prices.

This detailed visual analysis illustrates how intertwined sentiment analyses and market price movements are, and how the media impacts the market. It also shows how valuable sentiment analyses can potentially be for financial analyses and investment strategies.

ii. Correlation Between Sentiment and Bitcoin Prices:

To explore the relationship between sentiment scores and Bitcoin price movements, scatter plots and correlation matrices were used. These visualizations helped identify potential correlations and patterns, which are essential for developing predictive models. The scatter plots provided insights into the direct relationships between sentiment scores and price changes, while the correlation matrices quantified the strength and direction of these relationships. The analysis included both VADER and TextBlob sentiment analysis tools to compare the sentiment percentages (positive, negative, and neutral) against the daily Bitcoin price changes.

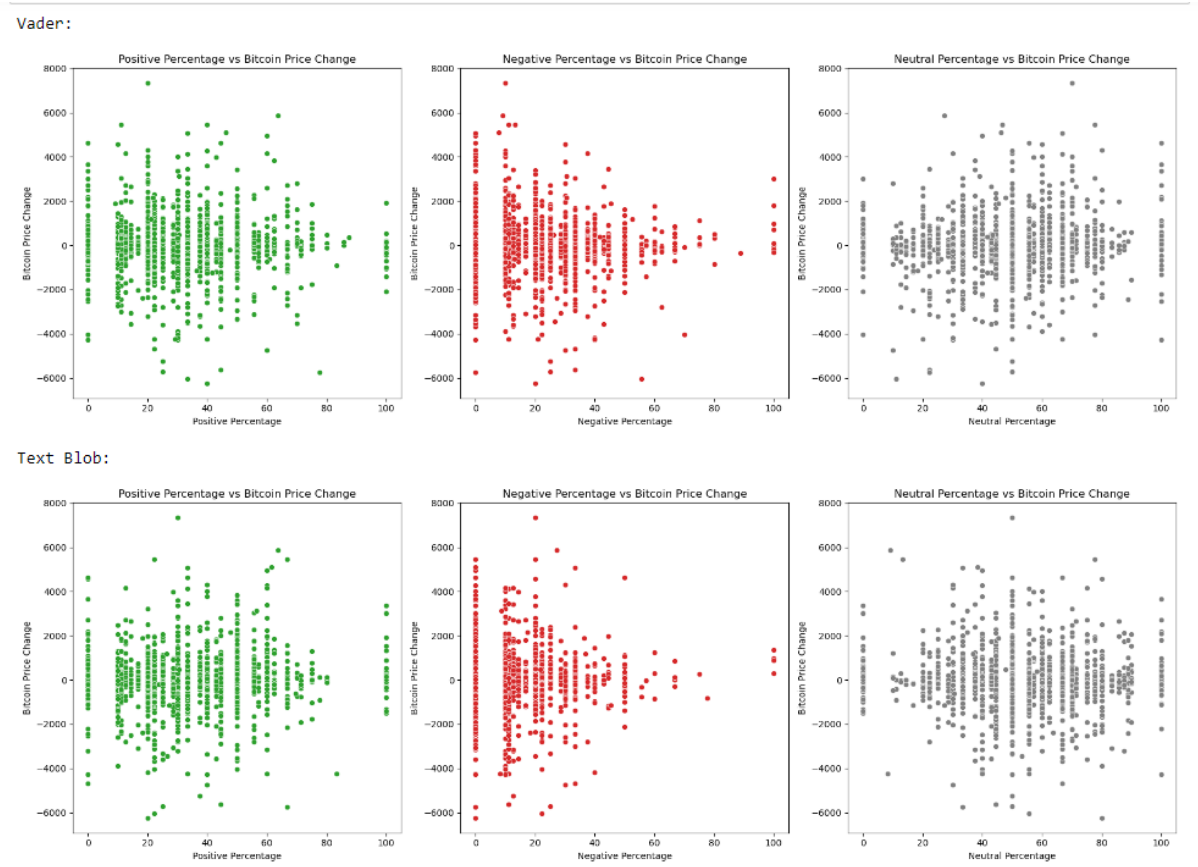

Figure 3: Comparison of Sentiment Analysis Results and Bitcoin Price Changes Using Vader and TextBlob.

- *Sentiment Analysis Scatter Plots*

The scatterplot in Figure 2 compares the sentiment analysis results obtained using Vader and TextBlob tools with Bitcoin price changes. According to the graph, a general increase in Bitcoin prices is observed as the percentage of positive sentiments increases. Both Vader and TextBlob tools show similar trends; however, Vader displays a more regular relationship between positive sentiments and price increases. In contrast, the results obtained with TextBlob present a more scattered distribution. Despite the distribution differences, high positive sentiment percentages in both cases are generally associated with significant price increases.

An increase in negative sentiments is shown to correlate with a decrease in Bitcoin prices. This relationship is more pronounced in the Vader graph, where an increase in negative sentiment percentages is consistently associated with noticeable price decreases. The TextBlob graph, while showing a similar trend, exhibits a less pronounced correlation between negative sentiments and price drops. Neutral sentiments do not exhibit a significant influence on Bitcoin price changes. Data from both tools indicate that as neutral sentiment percentages increase, the resulting price changes are generally low and variable, demonstrating that neutral sentiments do not significantly direct market movements.

The analysis presented in this scatterplot serves as a preliminary overview and does not conclude a preference for using Vader over TextBlob in the model development process. To further analyse, correlation matrices were investigated.

- *Correlation Matrices*

The data from the correlation matrices in Figure 4 illuminate the relationships between sentiment analysis conducted using Vader and TextBlob tools and Bitcoin price changes. Both positive and negative sentiment analyses and their impacts on Bitcoin price movements have been comparatively studied. Although statistically insignificant, a slight correlation was detected between positive sentiment percentages and Bitcoin price increases using the Vader tool ($r = 0.01$, $p > 0.05$). Similarly, a low level of negative correlation was observed between negative sentiment percentages and price decreases ($r = -0.14$, $p > 0.05$). However, TextBlob analyses have identified a more pronounced positive correlation between positive sentiments and price increases ($r = 0.12$, $p < 0.05$), and a stronger negative correlation between negative sentiments and price decreases ($r = -0.28$, $p < 0.01$).

These findings suggest that different sentiment analysis tools vary in their effectiveness at predicting price movements in the Bitcoin market. Despite the scatterplots initially suggesting a less clear distinction, the correlation matrix results reveal that TextBlob has demonstrated more accurate and pronounced impacts on price movements. Given these insights, it has been decided to employ TextBlob for model development in this context. This decision is informed by TextBlob's demonstrated ability to more accurately reflect market sentiments that significantly influence Bitcoin prices, which is crucial for developing robust market analyses and investment strategies.

iii. Distribution of Sentiment Scores:

Histograms and pie charts were utilized to display the distribution of sentiment scores obtained from both TextBlob and VADER analyses. These visualizations highlighted the frequency and proportion of various sentiment scores, offering insights into the overall sentiment landscape captured in the newspaper headlines. The histograms displayed the central tendencies and variability of sentiment scores, while the pie charts showed the relative distribution of positive, negative, and neutral sentiments. This comprehensive distribution analysis provided a clearer understanding of the sentiment dynamics and their occurrence within the dataset.

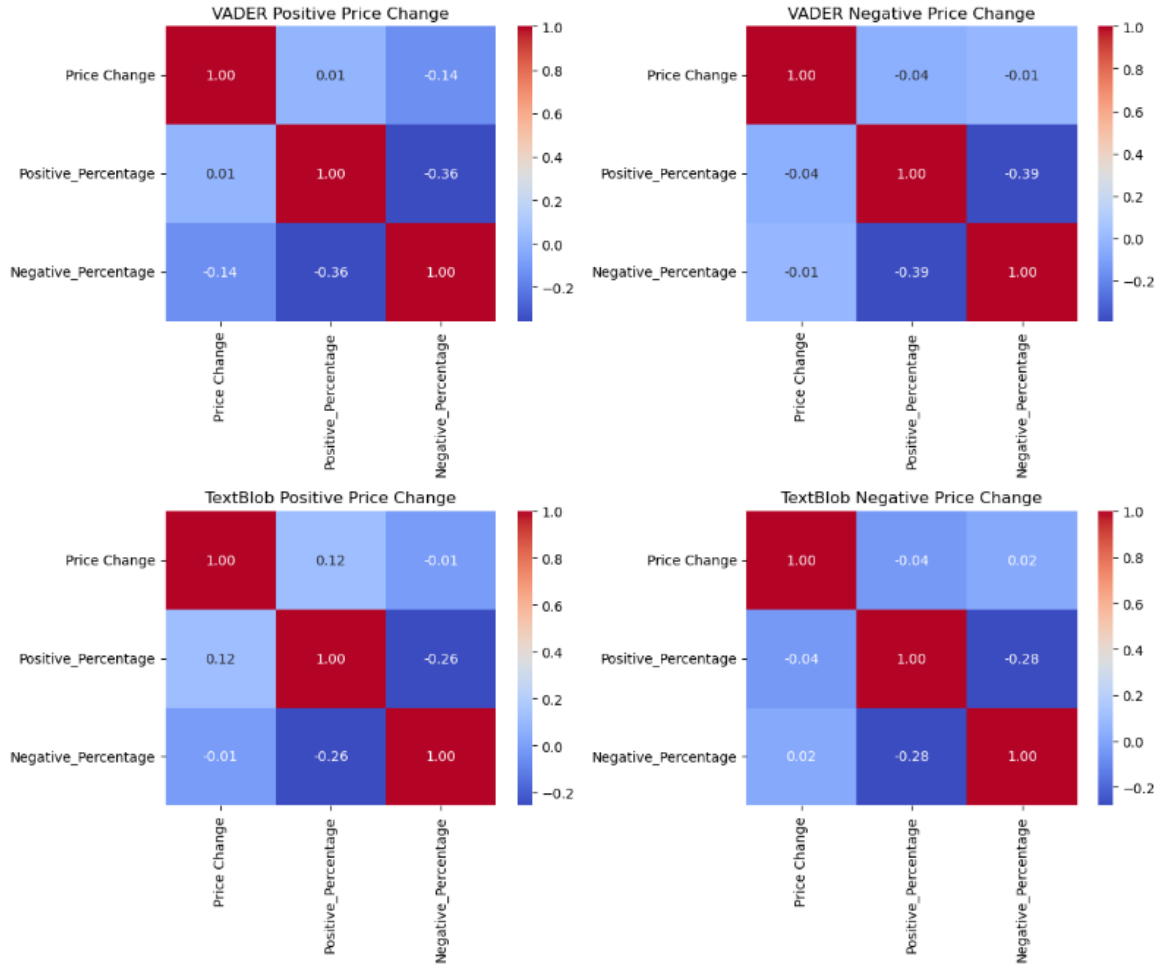

Figure 4: Correlation Matrices of Sentiment Scores and Bitcoin Price Changes for Vader and TextBlob.

- *Pie Charts*

The pie charts in Figure 4 display the distribution of sentiment analysis results conducted using the Vader and TextBlob tools. Initial observations indicate that the proportion of negative sentiment labels from TextBlob is significantly lower compared to Vader, while the proportion of neutral labels is higher. This suggests that Vader is more likely to assign sentiment scores to the same data or headlines compared to TextBlob.

Vader is specifically calibrated for sentiments expressed on social media and performs well with short texts, whereas TextBlob adopts a more general approach based on a pre-trained model (Sarkar, 2019: Hutto and Gilbert, 2014). The underlying mechanisms of both tools—Vader's rule-based approach and TextBlob's machine learning model—can lead to different interpretations of the same texts, especially when dealing with context-dependent or ambiguous sentiment expressions.

From these pie charts, it is not possible to determine which tool is more suitable for our project. TextBlob may be missing key sentiment words or may be accurately categorizing them as neutral. Moreover, these charts do not reveal the magnitude of sentiment; that is, while Vader may be scoring more headlines as negative, these might be mildly negative scores. Conversely, TextBlob might be scoring fewer headlines as negative but with more intensely negative scores. The results from both tools can be used to conduct in-depth semantic analyses on the provided texts to gain further insights. Further studies should be conducted to optimize the accuracy and comprehensiveness of these tools in sentiment analysis.

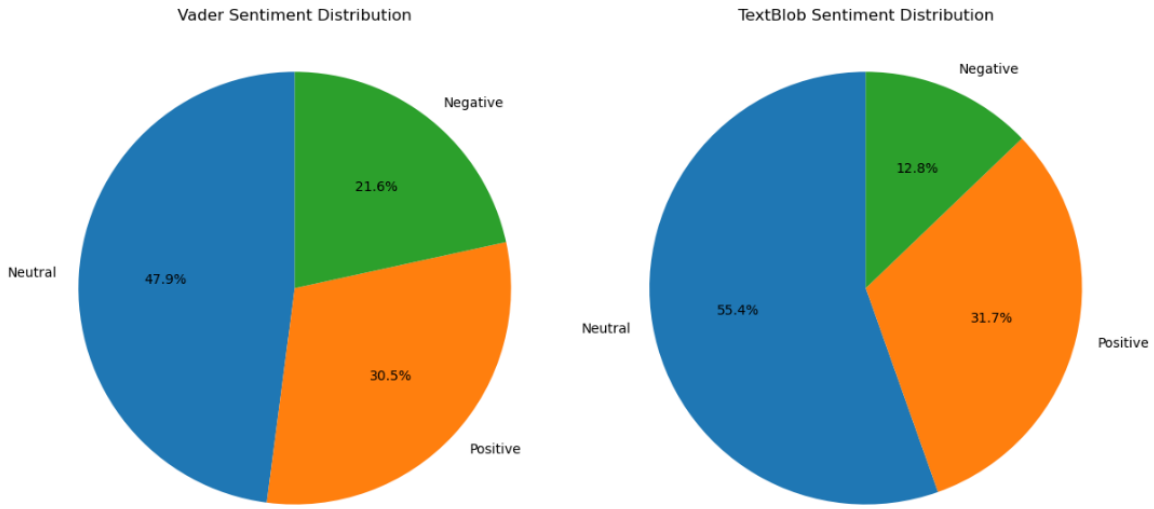

Figure 5: Sentiment Distribution Comparison Between Vader and TextBlob.

- *Histograms*

Histograms in Figure 5 display the distribution of sentiment scores obtained from texts analysed by the Vader and TextBlob tools. The most noticeable difference in the TextBlob histogram is the significantly higher bar between -0.04 and 0 compared to Vader. This suggests that TextBlob tends to classify certain headlines as neutral, whereas Vader might assess them as very slightly negative. The larger neutral bar in TextBlob could indicate a more conservative classification approach, potentially resulting in fewer false positives or false negatives, especially in texts containing ambiguous or complex sentiment expressions.

A second important observation is that Vader has a broader range on both the positive and negative sides compared to TextBlob. This indicates that Vader might capture emotional expressions in more detail and is more sensitive to emotional intensity. However, this broader range could also mean that Vader might give false alarms or assess mild emotional expressions as overly strong or weak.

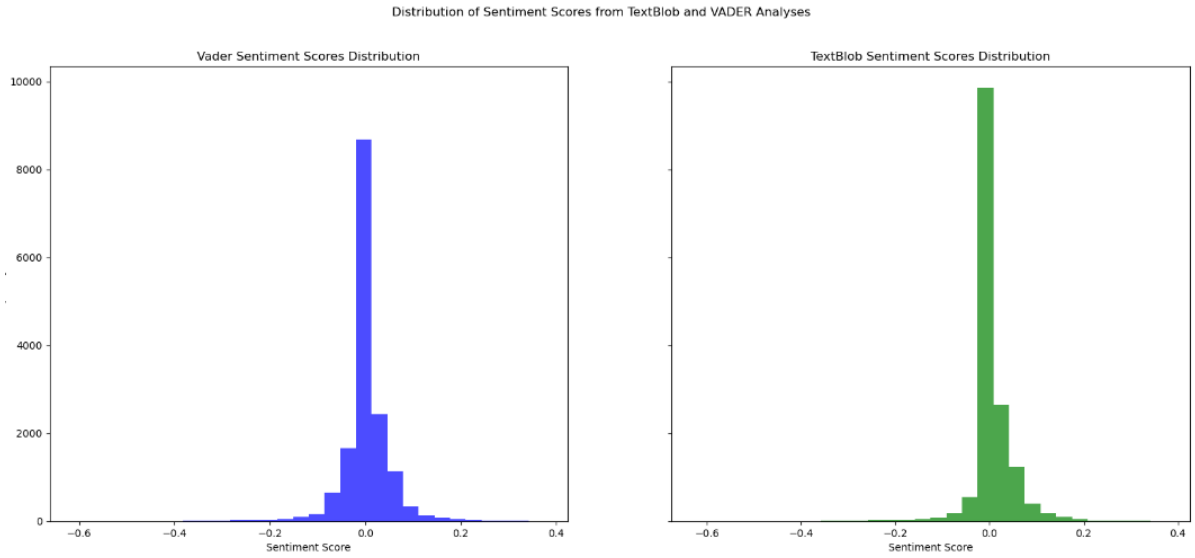

Figure 6: Distribution of Sentiment Scores from TextBlob and Vader Analyses.

These histograms provide valuable information on how each sentiment analysis tool perceives emotional tones. This information can be used to determine which tool might be more appropriate for specific project needs. For instance, Vader could be preferred in

situations requiring a broader spectrum of emotions, while TextBlob might be better suited for more balanced sentiment analysis.

The advantages and limitations of both tools should be considered when deciding on their use in sentiment analysis projects. These charts and related analyses offer a comprehensive look at the performance of sentiment analysis tools and how they handle different types of text.

iv. Word Cloud:

In this study, advanced text analysis techniques were used to examine the sentiment in Bitcoin-related headlines with the TextBlob sentiment analysis tool. Initially, data preprocessing included removing commonly recurring crypto-specific terms like 'Bitcoin' to avoid biased sentiment scores. However, terms directly indicating market performance, such as 'bull' and 'bearish', were kept due to their important role in sentiment interpretation.

To improve the clarity of the sentiment analysis, further preprocessing steps were taken before generating word cloud outputs. This included filtering the dataset to emphasize headlines with higher frequency, thus preventing the overrepresentation of outliers. This careful approach ensured that the sentiment analysis accurately reflected market sentiment, rather than being distorted by occasional peaks from less relevant news.

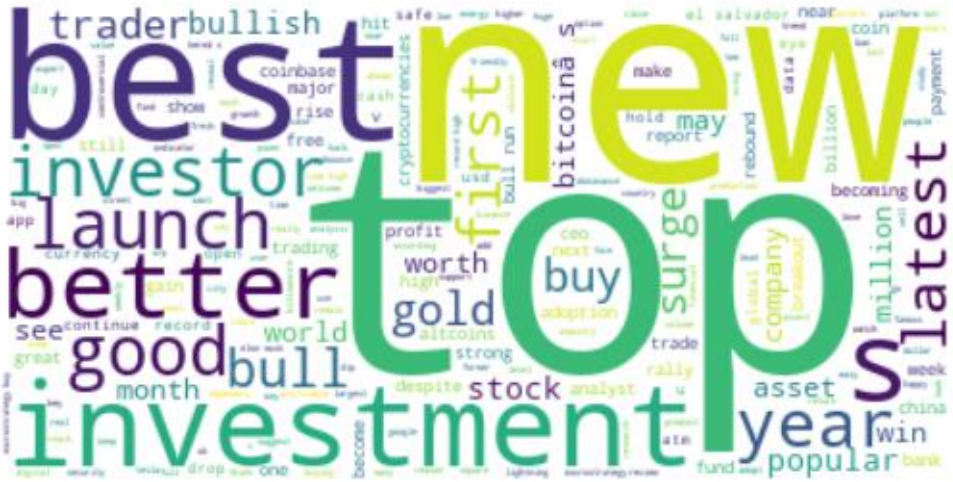

Figure 7: Word Cloud of Positive Sentiment Terms from TextBlob Analysis.

The word clouds generated from the refined data display the most frequent terms found in both highly positive and highly negative contexts, providing a visual representation of market sentiments. In the positive context, terms like "bullish," "launch," and "adoption" were notably common, indicating strong market optimism and confidence among investors (Figure 7). These terms are usually associated with periods of market growth and expansion, suggesting a positive outlook among market commentators.

On the other hand, the negative sentiment word cloud highlighted terms such as "crash," "scam," and "fraud," reflecting significant concerns about market stability and the integrity of market participants (Figure 8). These terms often appear during market downturns or when negative news related to fraud or regulatory crackdowns dominates the media.

These visualizations help in both understanding market sentiments qualitatively and supporting quantitative analyses by correlating specific terms with market movements. By analyzing the frequency and sentiment score of these terms, we can understand the emotional tone of the market and its potential impact on Bitcoin prices.

Additionally, this approach enables the creation of predictive models that use sentiment analysis to forecast market trends. Knowing how certain terms correlate with market

movements improves the accuracy of these models, offering valuable insights for investors and market analysts.

Figure 8: Word Cloud of Negative Sentiment Terms from TextBlob Analysis.

This comprehensive analysis, illustrated through word clouds and supported by data-driven insights, demonstrates the complex relationship between market sentiment and Bitcoin price dynamics. The findings underscore the importance of detailed sentiment analysis in understanding and predicting market behaviors, ultimately contributing to more informed investment strategies.

Model Development

Model development involved experimenting with several machine learning libraries to identify the most efficient one for predicting Bitcoin price fluctuations based on sentiment analysis. Various models were tested, and their performance was evaluated to ensure the selection of the most appropriate model for the study's objectives.

The development of machine learning models was carried out using three primary algorithms: Support Vector Machine (SVM), Logistic Regression, and Random Forest. The scikit-learn library was used to implement these models with its robust and user-friendly API. The dataset was split into training and testing sets to validate the performance of each model.

i. Training and Test Data

The dataset was divided into training and test sets using an 80:20 split ratio. This choice was made to ensure that the model had sufficient data to learn from while also retaining enough data to validate the model's performance. Furthermore, Kaabar (2023) explains that an 80:20 split is a commonly used practice in machine learning as it provides a good balance between training the model and evaluating its generalizability on unseen data. The training set was used to train the machine learning models, while the test set was reserved for evaluating the model's predictive performance.

ii. Feature Scaling

Feature scaling was applied to ensure that the data was normalized, which is particularly important for algorithms like SVM that are sensitive to the scale of the input features. The primary features used for training included sentiment scores and historical Bitcoin prices. Feature scaling was conducted using the StandardScaler from scikit-learn, which standardizes the features by removing the mean and scaling to unit variance. This process transforms the data to have a mean of zero and a standard deviation of one, ensuring that all features contribute equally to the model's performance and improving the convergence speed of gradient-based algorithms.

iii. Technical Indicators

Technical indicators were created using the TA-Lib library to provide additional features for the machine learning models. The TA-Lib library is a comprehensive technical analysis library that provides over 150 indicators for analysing financial market data. The following indicators were used in the study:

- *Moving Average Convergence Divergence (MACD):* According to Antonio, Alfredo, and Duque (2020), traditional technical analysis utilizes the Moving Average Convergence/Divergence indicator, which relies on exponential averages to emphasize the most current data in its calculations.
- *Relative Strength Index (RSI):* Gurrib and Kamalov (2019) highlight that the Relative Strength Index is a widely used technical indicator for assessing the rate of price changes, by comparing cumulative gains to cumulative losses.
- *Bollinger Bands (BB):* Gurrib and Kamalov (2019) highlight that the Bollinger Bands technique is grounded in the idea that a price significantly lower than its mean will revert to a normal level, suggesting a long position on the asset. According to Milstein et al., 2024, Bollinger Bands are composed of three bands: upper, middle, and lower. The middle band is a simple Moving Average, and the upper and lower bands are positioned at a distance from the middle band, which is determined by the standard deviation of the MA.
- *Exponential Moving Average (EMA):* The Exponential Moving Average is a key technical indicator used in stock analysis, functioning as a weighted version of the simple moving average. When the EMA line crosses, it generates buy or sell signals for the stock (Dhokane and Agarwal, 2018).
- *Simple Moving Average (SMA):* The Simple Moving Average calculates the average price of Bitcoin over a specified number of time periods (Karasu et al., 2020).

These indicators were added to the dataset and used as features in the machine learning models. The goal was to enhance the predictive power of the models by incorporating technical analysis insights along with sentiment scores, historical prices and daily volumes.

iv. Logistic Regression

According to Çolak (2023), logistic regression analysis allows for classification based on probability by estimating the dependent variable's values as probabilities. It utilizes three primary methods: binary, ordinal, and nominal logistic regression.

Evaluation metrics included accuracy, precision, recall, and F-1 score.

- Accuracy, a commonly used performance measure in both binary and multiclass classification problems, is defined as the proportion of correctly classified samples out of all samples (Zhou, 2021).
- Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positives. Generally, if you improve precision, recall may decrease, and if you improve recall, precision may decrease. Precision focuses on the accuracy of the positive predictions made, and recall focuses on capturing all relevant instances (Zhou, 2021).
- The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated using the formula:

$$F1 = \frac{2 \times P \times R}{P + R}$$

  where P is precision and R is recall. This metric is particularly useful when you need to balance precision and recall in your model evaluation (Zhou, 2021).
- Support refers to the number of true instances for each label in the dataset, indicating how many times each class appears in the output of the models used in this project.

v. Random Forest

Random Forest (RF), introduced by Breiman (2001), builds on Bagging (Bootstrap AGGregatING, a parallel ensemble learning method based on bootstrap sampling) by adding randomized feature selection. Instead of choosing the best split from all features, RF selects from a random subset, enhancing diversity and improving generalization. Despite its simplicity, RF often outperforms due to this added randomness and is more efficient to train compared to traditional Bagging methods (Breiman, 2001). The number of trees (n_estimators) and the maximum depth of the trees were tuned to optimize the model's performance. The evaluation metrics included accuracy, precision, recall, and F1-score.

vi. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised learning model used for classification tasks. It works by finding a hyperplane that best separates different classes in the feature space, maximizing the margin between the classes to improve generalization (Zhou, 2021). The SVM model was implemented to classify the Bitcoin price movements. GridSearchCV was used to determine whether to proceed with the linear or Radial Basis Function (RBF) kernel and to identify the optimal values for the regularization parameter (C) and the kernel coefficient (gamma). The model's performance was evaluated based on accuracy, precision, recall, and F1-score.

vii. Feature Engineering/ Expansion

Feature engineering, which involves deriving features from raw data and transforming them into formats suitable for machine learning models, is a vital step in the machine learning pipeline. This process is essential because well-engineered features can simplify the modelling task, leading to higher quality results from the machine learning pipeline (Zheng & Casari, 2018).

In this study, several technical indicators were implemented to improve the accuracy of Bitcoin price predictions. The main features used included sentiment scores from newspaper headlines and historical Bitcoin prices. To augment these features, several technical indicators incorporated using the TA-Lib library, which provides a comprehensive set of tools for technical analysis. The following indicators were added to the dataset:

1. *Moving Average Convergence Divergence (MACD):* This indicator helps identify changes in the strength, direction, momentum, and duration of a trend in Bitcoin prices. It uses exponential averages to emphasize the most current data (Antonio Alfredo and Duque, 2020).
2. *Relative Strength Index (RSI):* RSI measures the speed and change of price movements. It is used to identify overbought or oversold conditions in the market (Gurrib and Kamalov, 2019).
3. *Bollinger Bands:* These bands consist of three lines: the middle band, which is a simple moving average, and the upper and lower bands, which are standard deviations away from the middle band. Bollinger Bands help identify price volatility and potential overbought or oversold conditions Bollinger Bands help identify price volatility and potential overbought or oversold conditions (Milstein et al., 2024).
4. *Exponential Moving Average (EMA):* EMA places greater weight and significance on the most recent data points. This indicator helps identify the trend direction and potential reversals (Dhokane and Agarwal, 2018).
5. *Simple Moving Average (SMA):* SMA calculates the average price of Bitcoin over a specified number of time periods, providing insights into the overall trend direction (Karasu et al., 2020).

These indicators were calculated using the TA-Lib library in Python and added to the dataset to enhance the features used for training the machine learning models. Feature scaling was applied using the StandardScaler from the scikit-learn library to ensure all features

contributed equally to the model's performance. This standardization process, which transforms the data to have a mean of zero and a standard deviation of one, is particularly important for algorithms like SVM that are sensitive to the scale of the input features. Proper feature scaling ensures that all features are on a comparable scale, which significantly improves the performance and convergence speed of the SVM model (Baeldung, 2023; scikit-learn, 2023). Additionally, sentiment scores were adjusted using a custom lexicon specific to cryptocurrency. This lexicon included terms like "bullish," "bearish," and "hodl," which generic sentiment analysis tools might not capture effectively.

By incorporating these engineered features, the models were better equipped to understand and predict the complex dynamics of Bitcoin price movements. This comprehensive approach to feature engineering played a significant role in enhancing the model's predictive accuracy and robustness.

Discussion

i. Sentiment Analysis

Sentiment analysis for this project utilized two Python integrated sentiment analysis tools; TextBlob and VADER, to predict bitcoin prices evaluating the mood of Bitcoin-related newspaper headlines. By implementing these libraries, it aimed to ensure robustness and reliability in the sentiment analysis process. TextBlob and VADER were selected because they are strong at different things: TextBlob is strong at general text processing while VADER on the other hand is specifically designed for social media and informal text thus suited well for news heading analysis.

Hutto and Gilbert (2014) demonstrated that VADER was effective in analysing social media texts because it can handle informal expressions as well as fine-grained sentiments. They found that VADER performs better than other sentiment analysis tools with respect to F1 classification accuracy and generalizes well across domains. In their sentiment analysis of customer reviews Bonta et al. (2019) employed TextBlob explaining its efficacy and ease of use when it comes to deriving sentiment scores from text. TextBlob is well-known for general text processing and offers a complete set of tools for evaluating polarity and subjectivity. The use of TextBlob and VADER for sentiment analysis leveraged their strengths to ensure more accurate insights. Which made it possible to capture human emotions in text. The integration with these tools and our custom lexicon tailored to the cryptocurrency domain enhanced the analysis. These were inclusive of, “bullish”, “bearish”, and “hodl”. This domain-specific lexicon was integrated into TextBlob’s as well as VADER’s analyses thus ensuring precise sentiment classification through accounting for crypto jargon and expressions.

Textblob and VADER were both used to calculate sentiment scores. TextBlob provided sentiment polarities, which were adjusted using a custom lexicon that better reflected sentiments when using crypto related terms in headlines. In contrast, VADER produced compound sentiment scores indicating each headline as either positive or negative or neutral. That way both libraries are employed allowing for a comprehensive assessment by having the aggregated consistency of the sentiments that appear on both sources.

A simple aggregation was performed to represent the sentiment trend per day. (Figure 2) shows these daily groupings by positive, negative and neutral headings. Plotted the proportion of each sentiment type over time to determine how it could perceive Bitcoin's general mood of optimism or pessimism. Visualizations used were line plots, bar charts and histograms which depicted the trends and distribution of public news sentiment over time. Line plots showed weekly changes in average sentiments about Bitcoin from a positive viewpoint. However, visualizations were important for understanding what was happening with feelings. Visualization also allowed for observing the direct connection between different variables as

well as showing us their correlation force and direction. It revealed that this is not a straightforward correlation between prices and volumes

Analysis of sentiment scores and Bitcoin price movements was a key aim of this project. The data was processed using scatter plots and correlation matrices (Figures 3 and 4) to identify possible relationships and trends between changes in prices and sentiment scores. These visualizations gave an indication of some direct associations that were present while quantifying the strengths as well as the directions of such correlations. But this analysis also indicated how difficult it is to find stable trends amidst high frequency and highly variable data conditions.

A robust sentiment analysis framework was ensured by combining TextBlob with VADER, as well as a custom lexicon. However, issues were brought up regarding data collection and processing. It was based on headline news exclusively from English dailies which are limited; hence, there may be important global mood indicators not captured in the process. Nonetheless, given two sentiment analysis tools alongside a customized lexicon – it remains hard to fully capture cryptocurrency sentiment's intricacies.

Future research could include analysing a broader selection of news outlets for more comprehensive insights. Also, better language processing techniques for crypto-specific languages might be considered by employing more sophisticated techniques in natural language processing such as dependency parsing or semantic role labelling.

ii. Model Analysis

The model analysis in this study involved evaluating the predictive performance of various machine learning algorithms for Bitcoin price prediction using sentiment analysis of newspaper headlines. The three primary models tested were Support Vector Machine (SVM), Logistic Regression, and Random Forest. The evaluation metrics included accuracy, precision, recall, and F1-score.

The SVM model was used to classify Bitcoin price movements based on sentiment scores, historical price data and some technical indicators. SVM is particularly effective in high-dimensional spaces and operates by finding the optimal hyperplane that maximizes the margin between different classes in the feature space (Zhou, 2021). GridSearchCV was utilized to optimize the hyperparameters, including the regularization parameter (C) and kernel coefficient (gamma) for both linear and Radial Basis Function (RBF) kernels.

The SVM model's performance was evaluated on the test set and showed results, achieving a macro average accuracy of 36%, precision of 34%, recall of 36%, and an F1-score of 35%. Although these numbers might seem modest, they provide insights into the challenges of predicting Bitcoin price movements based on sentiment and technical indicators. These results indicate that while the SVM model has some predictive power, there is significant room for improvement, possibly through more sophisticated feature engineering or advanced modelling techniques.

Logistic regression estimates the relationship between the dependent variable and one or more independent variables by calculating probabilities using a logistic function (Çolak, 2023). According to Çolak (2023), logistic regression can be applied in three forms: binary, ordinal, and nominal logistic regression. The model's performance in this study was assessed using the same dataset, achieving a macro average of 35% accuracy, 32% precision, 34% recall, and 32% F1-score. These results suggest that while Logistic Regression is effective in some contexts, its performance may be limited compared to more complex models like SVM and Random Forest.

The Random Forest (RF) model, introduced by Breiman (2001), enhances the Bagging (Bootstrap Aggregating) method by incorporating randomized feature selection. Instead of selecting the best split from all features, RF chooses a random subset, which increases model diversity and improves generalization (Breiman, 2001). Despite its simplicity, RF often

outperforms other models due to this added randomness and is more efficient to train than traditional Bagging methods. The RF model in this study achieved a macro average of 31% accuracy, 29% precision, 30% recall, and 29% F1-score. The moderate performance of the Random Forest model suggests that while it captures some patterns in the data, there is potential for improvement, possibly through feature engineering or by exploring more advanced ensemble methods.

iii. Comparison of Model Performance

A comprehensive comparison of the three models revealed distinct differences in their performance. While the Support Vector Machine model exhibited the highest accuracy, Logistic Regression demonstrated superior precision and recall compared to the Random Forest model. The ensemble approach of Random Forest, leveraging multiple decision trees, provided a stable prediction framework, but did not achieve the same level of accuracy or precision as SVM or Logistic Regression in this study.

iv. Feature Importance and Technical Indicators

Incorporating technical indicators such as Moving Average Convergence Divergence, Relative Strength Index, Bollinger Bands, Exponential Moving Average, and Simple Moving Average alongside sentiment scores significantly enhanced the models' predictive power. The feature importance analysis indicated that combining sentiment scores with these technical indicators provided a comprehensive view of market dynamics. The SVM model leveraged these combined features effectively, demonstrating the value of integrating diverse data types to enhance predictive accuracy.

v. Trading Simulation

To evaluate the practical applicability of the model, a trading simulation was conducted comparing the performance of a single trader using the model's predictions against random trading decisions. The chosen trading model was the Support Vector Machine (SVM) due to its demonstrated accuracy. The simulation started with an initial capital of $10,000. The number of trades was determined by the length of the test set, which comprised 390 trades. For each trade, the trader could either strongly sell, sell, hold, buy or strongly buy, with the model-based strategy using the predicted labels. A separate random strategy was simulated for comparison, making random decisions for each trade.

The simulation function adjusted the capital based on the price changes and the trading decisions. For example, a strong sell would double the amount of capital influenced by the price change, while a hold would result in no change to the capital. This approach aimed to mimic real trading scenarios where different decisions impact capital differently.

The results showed that the average ending capital for the random strategy was $10,060.95, with a wide range from -$252,652.26 to $197,851.68. This significant variance indicates the unpredictable nature of random trading. In contrast, the model-based strategy showed a much higher and consistent outcome, with an ending capital of $123,100.40. This consistency suggests that the model-based strategy provided more stable and profitable results compared to random trading.

The results highlight the effectiveness of the model-based strategy in predicting Bitcoin price movements and making profitable trading decisions. The increase in ending capital for the model-based strategy underscores the potential of using machine learning models for financial predictions. This simulation demonstrates that while random trading can lead to extreme gains or losses, a model-based approach can provide more reliable and consistent returns. The consistency in the model-based strategy's outcomes also indicates that the model captures significant patterns in the data, which can be leveraged for more informed trading decisions.

vi. Future Work

To enhance the predictive capabilities of the models developed in this study, future work should focus on three key areas: incorporating advanced machine learning models, enhancing feature engineering and selection, and developing hybrid models and ensemble learning techniques.

Incorporating advanced machine learning models such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) can significantly improve prediction accuracy by capturing temporal dependencies and complex patterns in time-series data (Song et al., 2020, Nemes and Kiss, 2021).

Further refinement of feature engineering techniques, including the exploration of additional technical indicators, macroeconomic variables, and advanced sentiment analysis tools, can be important for improving model performance.

Developing hybrid models that combine the strengths of different algorithms, such as the interpretability of Logistic Regression, the complexity-capturing capabilities of SVMs, and the robustness of Random Forests, can lead to better predictive performance (Zhou, 2021). Ensemble learning techniques, like stacking and boosting, should also be explored to improve the model's generalization ability and robustness (Zhou, 2021).

By addressing these areas, future research can build on the findings of this study to develop more accurate and reliable models for Bitcoin price prediction, contributing to the growing field of financial forecasting and providing valuable tools for traders and investors in the cryptocurrency markets.

Conclusion

This study has meticulously analysed the interplay between media sentiment and Bitcoin price movements, leveraging advanced machine learning techniques and natural language processing tools. Through a robust analytical framework, the research examined the predictive capabilities of various models, including Support Vector Machines, Logistic Regression, and Random Forest, each providing unique insights into market dynamics.

A key takeaway from the research is the significant influence of media sentiment on market behaviour, emphasizing the need for sophisticated tools to parse and interpret complex datasets. The choice of TextBlob for sentiment analysis, following comparative evaluations, underscores the importance of selecting appropriate analytical tools that align with specific data characteristics and research objectives.

Moreover, the study highlights the potential of integrating technical indicators with sentiment data to enhance predictive accuracy. This dual approach not only enriched the analytical depth but also showcased the practical applicability through simulated trading scenarios, providing a realistic assessment of model performance in a volatile market environment.

In conclusion, while the models showed varied levels of efficacy, they collectively contribute to a deeper understanding of the factors driving cryptocurrency markets. The insights derived from this study are invaluable for developing refined predictive models and can serve as a foundation for further research into automated trading systems and market analysis strategies. This aligns with the ongoing evolution of financial technologies and the increasing reliance on data-driven decision-making in the digital economy.

References

Anamika, A. and Subramaniam, S. (2022) 'Impact of news sentiment on the cryptocurrency market'. *Applied Economics* [online]. Available from: https://doi.org/10.1080/00036846.2021.1950723 [Accessed 2 August 2024].

Antonio, A., Alfredo, R. and Duque, D. (2020) 'Machine learning applied in the stock market through the Moving Average Convergence Divergence (MACD) indicator'. *Investment Management and Financial Innovations* [online]. 17(4), pp. 44–60. Available from: https://www.proquest.com/docview/2477710440?pq-origsite=primo&sourcetype=Scholarly%20Journals [Accessed 25 July 2024].

Arjmand, M., Kazeminia, S. and Sajedi, H. (2024) 'Bitcoin price prediction based on financial data, technical indicators, and news headlines sentiment analysis using CNN and GRU deep learning algorithms', Third International Conference on Distributed Computing and High Performance Computing (DCHPC), Tehran, Iran, Islamic Republic of, 2024, pp. 1-7, doi: 10.1109/DCHPC60845.2024.10454082.

Badev, A.I. and Chen, M**.** (2014) *Bitcoin: Technical Background and Data Analysis*. Washington: Federal Reserve Board.

Bâra, A., Oprea, S. and Mirela, P. (2024) 'Insights into Bitcoin and energy nexus: A Bitcoin price prediction in bull and bear markets using a Meta-Model'. *Energy Reports* [online]. Available from: https://www.sciencedirect.com/journal/energy-reports [Accessed 2 July 2024].

BBC News (2021) 'Bitcoin: El Salvador makes cryptocurrency legal tender', *BBC News*, 9 June. Available at: https://www.bbc.com/news/world-latin-america-57398274 [Accessed: 2 August 2024].

BBC News (2024) 'Bitcoin: Crypto fans can now invest in exchange-traded funds - but what are they?'*BBC News*, 10 January. Available at: https://www.bbc.co.uk/news/technology-67916142#:~:text=The%20US%20has%20made%20the,pension%20funds%20to%20ordinary%20investors [Accessed: 2 August 2024].

Blanchard, A. (2007) 'Understanding and customizing stopword lists for enhanced patent mapping'. *World Patent Information* [online]. 29(3), pp. 236-244. Available from: https://www.worldpatentinformation.com/article/Understanding-and-customizing-stopword-lists-for-enhanced-patent-mapping [Accessed 2 July 2024].

Breiman, L. (2001) 'Random forests'. *Machine Learning*, 45(1), pp. 5-32. Available at: https://link.springer.com/article/10.1023/A:1010933404324 [Accessed 28 July 2024].

Chen, J. (2023) 'The implications of machine learning in the prediction of cryptocurrency prices'. *Journal of Financial Data Science*, [e-journal] 5(4), pp. 8-19. Available through: JSTOR database [Accessed 3 August 2024].

Cocco, L., Tonelli, R. and Marchesi, M. (2021) 'An agent-based artificial market model for studying the Bitcoin trading'. *IEEE Access*, 9, pp. 58576-58590.

Colak, Z. (2023) 'Predicting Financial Failure Using the Logistic Regression Model: Evidence from Istanbul Stock Exchange'. *International Journal of Economics, Business and Politics*, 7(1), pp. 184-202.

Colianni, S., Rosales, S. and Signorotti, M. (2015) 'Algorithmic trading of cryptocurrency based on Twitter sentiment analysis'. *CS229 Project*. Stanford University. Available at: https://cs229.stanford.edu/proj2015/028_report.pdf [Accessed 3 August 2024].

CoinGecko (2024) Glossary. Available from: https://www.coingecko.com/en/glossary [Accessed 01 July 2024].

CoinMarketCap (2024) Crypto Glossary. Available from: https://coinmarketcap.com/academy/glossary [Accessed 01 July 2024].

Mr.Creatonics (2023) *51 Cryptocurrency Glossary: Dictionary of Cryptocurrency and Bitcoin Terms*. CoinSutra. Available from: https://coinsutra.com/cryptocurrency-glossary-terminology/ [Accessed 01 July 2024].

CryptoNest (n.d.) Crypto Dictionary. Available from: https://cryptonest.co.uk/pages/crypto-dictionary [Accessed 01 July 2024].

Derakhshan, A. and Beigy, H. (2019) 'Sentiment analysis on stock social media for stock price movement prediction'. *Engineering Applications of Artificial Intelligence*, 85, pp. 569-578.

Dhokane, R. M. & Agarwal, S. (2024) 'Enhancing Stock Price Prediction with MACD and EMA Features Using LSTM Algorithm', in *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*. [Online]. 2024 IEEE. pp. 1–6. [Accessed 25 July 2024].

Fakharchian, S. (2023) 'Designing a forecasting assistant of the Bitcoin price based on deep learning using market sentiment analysis and multiple feature extraction'. *Soft Computing*, 27, pp. 18803-18827.

Foley, S., Karlsen, J.R. and Putniņš, T.J. (2019) 'Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed Through Cryptocurrencies?', *The Review of Financial Studies*, 32(5), pp. 1798-1853.

Franc, A. (n.d.) Topic Modeling and Network Analysis of Political News Articles. Available from: https://github.com/atfranc2/Topic-Modeling-and-Network-Analysis-of-Political-News-Articles [Accessed 2 August 2024].

Gurrib, I. and Kamalov, F. (2019) 'The implementation of an adjusted relative strength index model in foreign currency and energy markets of emerging and developed economies'. *Macroeconomics and Finance in Emerging Market Economies* [online]. 12(2), pp. 105–123. Available from: https://www-tandfonline-com.ezproxy.uwe.ac.uk/doi/epdf/10.1080/17520843.2019.1574852?needAccess=true [Accessed 25 May 2024].

Hossain, M.Z., Hossain, M.S. and Hashem, M.M.A. (2021) 'Sentiment analysis on social media data such as Twitter and Facebook using machine learning approaches: A

comprehensive review'. *IEEE Access*, [e-journal] 9, pp. 14937-14954. Available through: IEEE Xplore database [Accessed 3 August 2024].

Hossain, M.Z., Rahman, M.A., Islam, M.S. & Kar, S., (2021). 'Correlation of Sentiment Analysis between Tweets and Newspaper Headlines Using NLP'. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). IEEE, pp.135-139.

Huang, Q. (2015) Bitcoin price prediction based on fear & greed index. *SHS Web of Conferences*. [Online] 1812015-. Available at: https://doi.org/10.1051/shsconf/202418102015

Hutto, C.J. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text'. *Proceedings of the Eighth International AAAI Conference on Weblogs and SocialMedia* [online]. Available from: https://www.aaai.org [Accessed 03 August 2024].

Kaabar, S. & Chamberlain, M. (2024) Deep learning for finance: creating machine & deep learning models for trading in Python [online]. Available from: https://learning-oreilly-com.ezproxy.uwe.ac.uk/library/view/deep-learning-for/9781098148386/ch07.html [Accessed 25 July 2024].

Kamalov, F., Gurrib, I., and Rajab, K. (2024) 'Financial forecasting with machine learning: price vs return'. *Journal of Computer Science* [online]. Available from: https://doi.org/10.3844/jcssp.2021.251.264 [Accessed 2 July 2024].

Karasu, S., Altan, A., Bekiros, S. and Ahmad, W. (2020) 'A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series'. *Energy* [online]. 212, pp. 118750–118750. Available from: https://www-sciencedirect-com.ezproxy.uwe.ac.uk/science/article/pii/S0360544220318570?via%3Dihub [Accessed 25 July 2024].

Kean Chin, C. & Omar, N. (2020) Bitcoin Price Prediction Based on Sentiment of News Article and Market Data with LSTM Model. *Asia-Pacific Journal of information technology and multimedia*. [Online] 9 (1), 1–16.

Khyani, N., Lashari, S.A., Brohi, K. and Pallan, K. (2021) 'An Interpretation of Lemmatization and Stemming in Natural Language Processing'. In: *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE, pp. 1415-1419.

Kristoufek, L. (2015) 'What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis'. *PloS One*, 10(4), p. e0123923.

LexisNexis (n.d.) LexisNexis. Available from: https://www.lexisnexis.com/en-us/home.page [Accessed 01 July 2024].

Lison, P., Mowatt, D., and Tjong Kim Sang, E. (2019) 'Automatic detection of misleading content in news articles'. *Nature Machine Intelligence* [online]. 1(10), pp. 409-412. Available from: https://www.nature.com/articles/s42256-019-0112-6 [Accessed 2 July 2024].

Nakamoto, S. (2008) Bitcoin: A Peer-to-Peer Electronic Cash System. Available at: https://bitcoin.org/bitcoin.pdf [Accessed: 2 August 2024].

Nemes, L. and Kiss, A. (2021) 'Prediction of stock values changes using sentiment analysis of stock news headlines'. *Journal of Information and Telecommunication*, 5(3), pp. 375-394. Available from: https://www.tandfonline.com/doi/epdf/10.1080/24751839.2021.1874252?needAccess=true [Accessed 28 July 2024].

Nguyen Phuong, T., Nguyen Thanh, B. and Nguyen Hoang, M. (2024) 'Bitcoin price prediction using sentiment analysis'. *Journal of Big Data*, 11(1), pp. 1-15.

Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B., 2016, October. Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

Pant, D.R., Neupane, P., Poudel, A., Pokhrel, A.K. and Lama, B.K., 2018, October. Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS) (pp. 128-132). IEEE.

Ranjan, S., Kayal, P. and Saraf, M. (2023) 'Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach'. *Computational Economics*, 61, pp. 1617–1636.

Rathore, S., Ahmad, A., Paul, A. and Rho, S., 2022. The role of big data analytics in Internet of Things. *Computer Communications*, [e-journal] 119, pp. 137-143. Available at: https://doi.org/10.1016/j.comcom.2017.12.007 [Accessed 3 August 2024].

Roni, E. F., Alit, R., and Buana, R.C. (2023) 'Sentiment Analysis of Crypto Coin on Twitter Data Using Text Mining Method with K-Means Clustering Case Study: Bitcoin, Ethereum, and Binance'. In: *2023 IEEE 9th Information Technology International Seminar (ITIS)*. [Online]. 2023 IEEE. pp. 1–6.

Sarica, S., & Luo, J. (2021). 'Stopwords in technical language processing'. *PLoS ONE*, 16(8)

Sarkar, D. (2019) *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing* [online]. New York: Apress. Available from: https://doi.org/10.1007/978-1-4842-4354-1 [Accessed 03 August 2024].

Sattarov, O., Jeon, H.J., Oh, R. and Lee, J.D. (2020) 'Forecasting bitcoin price trends based on big data analysis'. In: *Big Data Analytics for Cyber-Physical Systems*. Elsevier, pp. 167-187.

Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L. and Cheng, Z., 2020. 'Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model'. *Journal of Petroleum Science and Engineering*, 186, p.106682. Available from: https://www-sciencedirect-com.ezproxy.uwe.ac.uk/science/article/pii/S0920410519311039?via%3Dihub [Accessed 28 July 2024].

Vo, N.N., Nguyen, Q.C. & Ock, C.Y., 2019. Cryptocurrency price prediction using news sentiment and technical indicators. In: 2019 International Conference on Advanced Computing and Applications (ACOMP). IEEE, pp.78-83.

Ye, Z., Shi, Y. & Wu, B. (2022) Bitcoin price prediction using sentiment analysis of social media comments. *IEEE Access*, 10, pp. 39757-39767.

Yermack, D. (2015) 'Is Bitcoin a Real Currency? An Economic Appraisal', in *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data*, pp. 31-43.

Zhou, Z.-H. (2021) *Machine Learning*. SpringerLink [online]. Available from: https://link-springer-com.ezproxy.uwe.ac.uk/book/10.1007/978-981-15-1967-3 [Accessed 28 July 2024].