**AN ANALYSIS OF THE INFLUENCE OF NEWS SENTIMENT ON CRYPTOCURRENCY PRICES**

**MUHAMMAD ZAHRIEL BIN ISMAIL**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY**

**UNIVERSITY OF MALAYA**

**kuala lumpur**

**2023**

**UNIVERSITY OF MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: Muhammad Zahriel Bin Ismail (I.C/Passport No: 970520-56-5345)

Matric No: 22085509

Name of Degree: Master of Data Science

Title of Project Thesis: Validating the Predictive Power of News on Cryptocurrency Prices Using Machine learning

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;

(2) This Work is original;

(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;

(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;

(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya (“UM”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;

(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate’s Signature Date:

Subscribed and solemnly declared before,

Witness’s Signature Date:

Name:

Designation:

# Abstract

<Placeholder>

# Acknowledgments

<Placeholder>

Table of Contents

[Abstract 3](#_Toc174345842)

[Acknowledgments 4](#_Toc174345843)

[1.0 Introduction 10](#_Toc174345844)

[1.1 Background Study 10](#_Toc174345845)

[1.2 Problem Statement 11](#_Toc174345846)

[1.3 Research Objectives 12](#_Toc174345847)

[1.4 Research Significance 12](#_Toc174345848)

[1.4.1 Enhanced Market Understanding 12](#_Toc174345849)

[1.4.2 Identification of Optimal Predictive Techniques 13](#_Toc174345850)

[1.4.3 Correlation and Impact Analysis 13](#_Toc174345851)

[1.4.4 Risk Management 13](#_Toc174345852)

[1.4.5 Contribution to Academic Knowledge 13](#_Toc174345853)

[1.5 Research Motivation 14](#_Toc174345854)

[1.5.1 Addressing Market Volatility 14](#_Toc174345855)

[1.5.2 Enhance Predictive Models 14](#_Toc174345856)

[1.5.3 Contribution to Financial Technologies 14](#_Toc174345857)

[2.0 Literature Review 15](#_Toc174345858)

[2.1 Cryptocurrency Price Predictions 15](#_Toc174345859)

[2.2 News Platforms Influence in Cryptocurrency 16](#_Toc174345860)

[2.3 Cross Validation of Existing Models and Techniques 19](#_Toc174345861)

[2.4 Critical Analysis 21](#_Toc174345862)

[2.5 Literature Summary 25](#_Toc174345863)

[3.0 Methodology 27](#_Toc174345864)

[3.1 Obtain Phase 29](#_Toc174345865)

[3.1.1 News Data Collection 29](#_Toc174345866)

[3.1.1.1 Article Indexing 30](#_Toc174345867)

[3.1.1.2 Article Scrapping 31](#_Toc174345868)

[3.1.2 Historical Price Data Collection 32](#_Toc174345869)

[3.1.3 Trade Volume Data Collection 32](#_Toc174345870)

[3.2 Scrub Phase 32](#_Toc174345871)

[3.2.1 News Data 32](#_Toc174345872)

[3.2.2 Price Historical Data 33](#_Toc174345873)

[3.2.3 Trade Volume 33](#_Toc174345874)

[3.3 Explore Phase 33](#_Toc174345875)

[3.3.1 Sentiment Analysis Data 33](#_Toc174345876)

[3.3.2 Historical Prices and Trade Volume 33](#_Toc174345877)

[3.4 Modelling Phase 34](#_Toc174345878)

[3.4.1 Model Selection 34](#_Toc174345879)

[3.5 iNterpret Phase 35](#_Toc174345880)

[3.5.1 Sentiment Analysis Modelling 35](#_Toc174345881)

[3.5.2 Price Prediction 35](#_Toc174345882)

[5.0 Results and Discussions 36](#_Toc174345883)

[5.1 Comparative Analysis 36](#_Toc174345884)

[5.2 Hypothesis Validation 36](#_Toc174345885)

[6.0 Deployment 37](#_Toc174345886)

[7.0 Reproducible Study 38](#_Toc174345887)

[8.0 Conclusion 39](#_Toc174345888)

[References 40](#_Toc174345889)

List of tables

[Table 1.3 Research Questions and Objectives 11](#_Toc173168945)

[Table 2.1 Cryptocurrency Segmentation……………………………………………………..15](#_Toc173168460)

[Table 2.2 News Platform Findings 17](#_Toc173168461)

List of Figures

[Figure 1.0 Proposed Pipeline 26](#_Toc173696692)

# Introduction

This section focuses on the background analysis, defined problem statements and research goals and scope of project.

## 1.1 Background Study

In recent years, cryptocurrencies have emerged as a significant innovation in the financial sector, offering decentralized alternatives to traditional fiat currencies. By utilizing blockchain technology, cryptocurrencies facilitate secure, peer-to-peer transactions internationally without the need for intermediaries like banks. Since the advent of Bitcoin, thousands of cryptocurrencies have been developed with differing features and applications.

Due to the differing nature of cryptocurrency as compared to traditional financial tools (currencies, assets, etc.) it is influenced by several different factors, including market demand, technological advancements, regulatory news and media sentiment. News platforms play a significant role in conveying public perception and investor behavior towards cryptocurrencies as these platforms disseminate information regarding cryptocurrencies such as regulatory changes, market trends and expert opinions which in turn can significantly influences investor sentiment (Bowden et al., 2022).

Research has shown that media sentiment, particularly from reputable news sources, can influence cryptocurrency prices. Positive news can drive up prices by increasing investor confidence and demand, while negative news can lead to price drops by causing panic and sell-offs. This interplay between news and cryptocurrency prices underscores the importance of sentiment analysis in understanding and predicting market movements (Sapkota, 2022).

In this context, analyzing the sentiment of news articles from news platforms like CoinDesk can provide valuable insights into future price trends of cryptocurrencies. By leveraging machine learning techniques to assess the sentiment conveyed in news articles, researchers and investors can develop predictive models to forecast price movements and make informed investment decisions (Chen et al., 2023).

Further understanding the differences between traditional stock markets and cryptocurrency markets is essential for grasping the unique dynamics at play within these two financial realms. Traditional stock markets are typically driven by fundamental factors such as company earnings, economic indicators, and broader market trends, which provide a more stable and predictable environment for investors. In contrast, cryptocurrencies represent a purely speculative market, characterized by their extreme volatility and the significant influence of market sentiments.

As a result, identifying and understanding the key themes and potential keywords related to cryptocurrency markets becomes increasingly important. These themes and keywords, which may include terms associated with market sentiment, regulatory changes, technological developments, or broader economic conditions, are crucial for generating deeper insights into how the cryptocurrency market operates (Aysan, Caporin, & Cepni, 2024).

## 1.2 Problem Statement

The cryptocurrency market, characterized by its volatility and lack of regulatory oversight, presents significant challenges for investors and market analysts. Unlike traditional financial markets, the value of cryptocurrencies is heavily influenced by public perception and sentiment, often driven by news articles and social media discussions. While traditional financial models struggle to account for the rapid and unpredictable changes in cryptocurrency prices, sentiment analysis offers a promising approach to understanding these dynamics. However, there is a lack of comprehensive studies examining the impact of news sentiment on cryptocurrency prices using advanced sentiment analysis techniques as well as highlighting sentiment decay and the lingering effects of both positive/negative sentiments. Previous studies have demonstrated that models incorporating sentiment analysis outperform those relying solely on historical price data. For instance, a study evaluating cryptocurrency price prediction using sentiment analysis and Long Short-Term Memory (LSTM) networks showed that integrating sentiment information from news articles significantly improved prediction accuracy compared to models using only historical prices (Anh-Dung et al., 2019).

This study aims to fill this gap by investigating the relationship between news sentiment from reputable news sources and cryptocurrency prices from a differing perspective.

## 1.3 Research Objectives

The following table contains the research questions and research objectives of this study. The table is as follows:

Table 1.3 Research Questions and Objectives

|  |  |
| --- | --- |
| Research Question | Research Objective |
| How does the sentiment of news articles correlate with daily price changes in cryptocurrencies? | Investigate the influence of news sentiment on cryptocurrency prices. |
| How can sentiment scores from news articles be integrated with historical price data to improve prediction accuracy | Develop a model for predicting cryptocurrency price trends using sentiment analysis. |
| What are the limitations of using sentiment analysis for predicting cryptocurrency prices, and how can these be mitigated | Assess the reliability of sentiment analysis in the context of cryptocurrency markets. |
| How reliant are cryptocurrencies on the performances of major market tokens. | Explore the relationship between tokens and large tokens to provide context on how heavily tokens can be influenced by different tokens. |

## 1.4 Research Significance

This section focuses on the potential advantages and impact of this study.

### 1.4.1 Enhanced Market Understanding

Generating insight by exploring the relationship between news sentiments and cryptocurrency prices will assist in enhancing the understanding of the cryptocurrency market. This helps arm decision makers and stakeholders with more information when handling cryptocurrency.

### 1.4.2 Identification of Optimal Predictive Techniques

By comparing multiple models, the research can identify which model provides the most accurate predictions for cryptocurrency prices based on sentiment analysis. This can guide investors, traders, and financial analysts in selecting the most effective predictive tools for their needs.

### 1.4.3 Correlation and Impact Analysis

The study can analyze the correlation and impact of news sentiments on cryptocurrency, this will assist in understanding how differing currencies are affected and to what extent it can be expected to be affected. This also includes the potential time lagged effect and decay of sentiments.

### 1.4.4 Risk Management

By understanding the sentiment and potential future price movements, stakeholders can better understand and manage the risk when involved with the cryptocurrency market. This study helps highlight and potentially assess potential risks in the market based on certain criteria.

### 1.4.5 Contribution to Academic Knowledge

The research conducted can contribute to academic literature by providing empirical evidence on the comparative performance of different models in the context of cryptocurrency and sentiment analysis. This can serve as a foundation for future research, encouraging further exploration and refinement of predictive models in financial studies.

## 1.5 Research Motivation

This section covers the research motivations behind the thesis and the importance/relevance of the research.

### 1.5.1 Addressing Market Volatility

Cryptocurrencies are known for their high volatility, which poses significant risks and opportunities for investors. Understanding how news sentiment influences cryptocurrency prices can provide insights into market dynamics, helping investors to make more informed decisions and potentially reducing the financial risks associated with trading these assets.

### 1.5.2 Enhance Predictive Models

Existing models for predicting cryptocurrency prices often focus on historical price data and technical indicators. By incorporating sentiment analysis of news articles, this research aims to enhance the accuracy of predictive models. This can lead to the development of more robust tools for financial forecasting, benefiting traders, analysts, and financial institutions.

### 1.5.3 Contribution to Financial Technologies

The integration of sentiment analysis into cryptocurrency market predictions represents a novel application of natural language processing (NLP) and machine learning in the field of FinTech. This research can contribute to the advancement of FinTech by demonstrating how innovative techniques can be applied to improve market predictions, ultimately fostering innovation and technological growth in the financial sector.

# 2.0 Literature Review

This section focuses on providing a brief understanding of the several aspects of studies that have contributed to the subject of this study or studies that can be closely attributed to the context.

## 2.1 Cryptocurrency Price Predictions

This section focuses on previous works that have conducted cryptocurrency price predictions and the methodologies used to derive their conclusions. Cryptocurrencies operate on a peer-to-peer-based transaction system secured through cryptographic algorithms. Their prices are notably volatile and have become significant investment assets. The high volatility and dependency on other cryptocurrencies necessitate robust forecasting models to predict price movements accurately. Previous research in predicting cryptocurrency price movements through sentiment analysis (X formerly known as Twitter) and supplementary cryptocurrency data has shown that cryptocurrency prices are also highly co-dependent on other cryptocurrencies within the same network (Ethereum, TRC20, etc.) (Parekh et al., 2022). The research also presents that deep learning models have better overall performance in predicting cryptocurrency prices as compared to traditional systems.

Similarly conducted research has also emphasized the importance of market sentiment derived from social media platforms have a significant impact on the daily prices of cryptocurrencies. The research explored three segments of tokens, established, emerging and meme tokens. This study identified that the prices of emerging tokens were more sensitive to market sentiment whereas established and meme tokens tend to be less affected (Koltun & Yamshchikov, 2023). The tokens utilized in the study are commonly defined as the following:

Table 2.1 Cryptocurrency Segmentation

|  |  |  |  |
| --- | --- | --- | --- |
| Token Type | Definition | Characteristic | Examples |
| Established | Cryptocurrencies with long history and significant market capitalization | * Well established in the market * High Market Capitalization * Greater acceptance and adoption | Bitcoin (BTC), Ethereum (ETH) |
| Emerging | Newer Cryptocurrencies that have potential for growth | * Relatively new in the market * High Volatility * Growing acceptance and adoption | Polkadot (DOT),  Chainlink (LINK) |
| Meme | Cryptocurrencies created as a joke with little to no underlying value | * Created as a joke * Low Market Capitalization * Large Price Swings based on Trends | Shiba Inu (SHIB),  Dogecoin (DOGE) |

This does not include cryptocurrencies such as “stable” coins which are directly influenced by specific economic factors such as Currency values or object standards.

## 2.2 News Platforms Influence in Cryptocurrency

This section focuses on the different processes conducted by similar studies and comparing the insights provided by each technique to identify the most suitable method for this study. A previous study that had processed data from major news providers such as Wall Street Journal utilized filters to filter the news based on a score-based system such as the relevance of the news to the identified cryptocurrency and the novelty of the news (Importance of this news within a 24-hour window). The study also split the sentiment based on the headline and the text body of the article, this allows for different granular levels of sentiment analysis (Rognone, Hyde, & Zhang, 2022).

Furthermore, a different study exploring the signal theory between covid-19 news/headlines and the relations to cryptocurrency markets identified a significant non-linear relationship between the news sentiments and cryptocurrency returns. This suggest that traditional linear models may not be able to fully capture the relationship between the two, this non-linear relationship was identified by utilizing transfer entropy to measure the causality between the news sentiment and cryptocurrency prices. The study also identifies that news with higher entropy transfer values have stronger impacts on the prices of cryptocurrencies, suggesting that certain types of news have higher significance than the rest (Banerjee et al., 2022).

Similarly, another study that formulated the “Cryptocurrency Uncertainty Index” explores the relationship between news sentiment and cryptocurrencies. It identifies that major events that have direct effects on market activities such as policy changes increase the price uncertainty of cryptocurrencies. These include events such as political events (Brexit, U.S Elections), regulation changes (Initial Coin Offering (ICO) banning in certain countries such as China) , Technological advancements (Launch of Defi Technology) and Cyberattacks (Hacking of large cryptocurrency platforms) have large influence in the uncertainty of the coins (Lucey, Vigne, Yarovaya, & Wang, 2022).

The following table summarize the findings in this section and the key takeaways provided by the sources:

Table 2.2 News Platform Findings

|  |  |  |  |
| --- | --- | --- | --- |
| **Study** | **Data Source** | **Sentiment Analysis Method** | **Key Finding** |
| [1] | Ravenpack | Utilize Ravenpack news indices (Data is already processed) | The granularity of the news is important |
| [2] | Ravenpack Database, Coinbase | Utilize Ravenpack news indices (Data is already processed) | Specific events regardless of the sentiment score has higher impact as compared to other events. Identifies the relationship as non-linear between news and prices |
| [3] | LexisNexis Business Database | N/A (Not mentioned/specified in the research journal) | Price of Cryptocurrencies are heavily influenced by activities that have direct effect on market activities. |

Based on the table, the following aspects need to be explored within this study to improve the understanding of the relationship between news sentiment and cryptocurrency prices:

1. Highlighting key events: News that has significant effects on market activities needs to be identified.
2. Sentiment Analysis Methods: Due to the studies utilizing pre analyzed sources, exploring more nuanced methods may provide better insight into the relationship between news sentiment and cryptocurrency prices.
3. Define the granularity of the data: Exploring the need to separate the headlines and body of the news source may be required to understand the influence of the news at different granularities and effects on the cryptocurrency prices.

## 2.3 Cross Validation of Existing Models and Techniques

This section covers the preprocessing techniques and modelling used in the context of news and/or cryptocurrency data. The following table identifies the different studies and the techniques utilized:

|  |  |  |
| --- | --- | --- |
| Study | Data Source | Technique Utilized |
| (Parekh et al., 2022) | X (Formerly known as Twitter) | Normalize the Prices towards the Maximum Price (0-10), utilize VADER (Valence Aware Dictionary and sEntiment Reasoner) for sentiment analysis. |
| (Gadi & Sicilia, 2024) | Cryptolin Corpus (2683 articles related to cryptocurrency) | Utilized several sentiment analysis models:   * Vader * Textblob * Flair * FinBert   FinBert had the best performance among the models utilized in the context of financial news. |
| (Sapkota, 2022) | LexisNexis Database, Investing.com  (17,490 News articles) | Extraction of 3 key data points from the articles (Headline, Publisher Name and Publish Date).  Sentiment Analysis based on four dictionaries:   * Harvard-IV general-purpose psychological dictionary (GI) * Quantitative Discourse Analysis Package (QDAP) dictionary * Henry's (2008) finance-specific dictionary (HE) * Loughran and McDonald (2011) finance-specific dictionary (LM)   Aggregate all news on a day-to-day basis. |

Based on this table we can identify that finance related news sentiments need to be tested with multiple sentiment analysis models. As Identified by the study [2], FinBert showed the best performance with finance news, exploring the performance of ‘Vader’ as compared to ‘FinBert’ needs to be explored during the sentiment analysis phase to identify major differences in more nuanced data. To ensure data is easier to process, normalizing the prices by max price would ensure easier data processing.

## 2.4 Critical Analysis

This section covers the summation of the previous sections as well as providing critical analysis on the literature to further support this study. The following is the critical analysis table:

|  |  |  |  |
| --- | --- | --- | --- |
| Study | Dataset | Findings and Results | Gaps |
| (Sapkota, 2022) | Daily Price Movements and Bitcoin related news from major English News Platforms | Identified that financial sentiment has significant long-term impact and that psychological sentiment were insignificant in predicting volatility.  Positive sentiment news increased the volatility of Bitcoin and potentially identifies that optimism is a primary driver. | Utilization of generic sentiment dictionaries may not have caught the finance nuances of the text.  Data is sourced from general news platforms which may not capture the sentiment or insight from more crypto-centric platforms. |
| (Rognone, Hyde, & Zhang, 2020) | Daily Bitcoin Trade prices and volume data.  Bitcoin related news and forex foreign exchange rate prices. | Supports the conclusion that only positive news has immediate effects on Bitcoin returns with negative news having a delayed positive relationship which differs from traditional currencies.  The study also acknowledges bubble periods which are periods of strong growth for Bitcoin where positive news has increased significance on the prices. | Only focuses on the context of one cryptocurrency (BTC) and findings may or may not be applicable to other currencies.  The study does not explore the potential of a non-linear relationship between the sentiment and Bitcoin prices. |
| (Koltun & Yamshchikov, 2023 | X (formerly known as Twitter) tweet data and pricing information of 12 different Cryptocurrencies. | The study identified that emerging tokens were sensitive to public sentiment while more established tokens were less influenced.  It also highlights that there is a dependance of the size of the cryptocurrency and the susceptibility of it to public sentiment where smaller cryptocurrencies are more influenced and suitable for speculation via sentiment analysis. | Due to the usage of Social Media data to generate sentiment, the data may be susceptible to noise generated by biased sources which may over or under exaggerate the overall sentiment. |
| (Aysan, Caporin, & Cepni, 2024) | Intraday Price data for 14 different cryptocurrencies and news articles from “Thomson Reuters Marketpscyh Analytics” | The study acknowledges that specific news themes have higher effects on cryptocurrency prices as compared to other related news.  It also highlights that larger coins such as Bitcoin have a direct correlation to other Bitcoin network coins such that if Bitcoin prices increase, other coins also undergo an increase. | The study only utilizes one news source to generate its sentiment and does not explore the potential for sentiment decay of the sentiment scores, this could potentially identify the lasting effects of sentiments. |
| (Sajter & Kuzmic, 2024) | Cryptocurrency price index and Covid-19 related news from three different news databases | The study similarly highlighted that specific types of news have higher significance on different cryptocurrencies.  It also identified that the cryptocurrency market displayed a low negative influence on returns in relation to the panic index (Levels of panic/fear in the general market) which highlights the stability of the crypto market in periods of high panic. | The study explores only established cryptocurrencies in the market and does not explore the significance of news against emerging or other classifications of cryptocurrencies.  The study also utilized only U.S based news sources which may make the findings not applicable to other regions. |
| (Helmi, Çatık, & Akdeniz, 2023) | Central Bank Digital Currency (CBDC) News indices, Cryptocurrency (Bitcoin) Uncertainty Measures and Market Value | News related to CBDC increased uncertainty index of Cryptocurrency market value.  Positive CBDC news correlates to negative impacts in Bitcoin Prices.  The impact of CBDC news increases during specific periods of rapid development of CBDC such as during COVID-19. | The study utilized only one model to conduct the analysis, expanding this to more models could assist in validating the results of the study.  Granularity of study is identified as to be analyzed on a weekly basis, reducing this to a daily basis may identify more information. |
| (Chokor & Alfieri, 2021) | Daily Price Movements of Specific Cryptocurrencies, News within Market Event Windows (3-days within events) | Prices of Cryptocurrencies tend to react negatively towards events that increase market regulations. This identifies that investors prefer a more unregulated market. The effect of regulatory news does not have significant effect on cryptocurrencies that already have privacy regulations. | The study does not evaluate more recent regulations and does not differentiate between large and small regulatory changes. By identifying the type of news, it could identify the potential impact of the news. |
| (Coulter, 2022) | News Articles related to Bitcoin, Daily Cryptocurrency Prices | The study identified that news related to crime and financial governance had negative effects on Bitcoin.  It also highlights that the geographical source of the news may influence the impact of the news against the price of Bitcoin | The study only analyzed a single cryptocurrency (BTC) against the news sentiment which may not be applicable to other cryptocurrencies. It also highlighted that the usage of different news sources is needed for validation of the hypothesis. |

Based on the critical analysis, it highlights that Bitcoin is the primary cryptocurrency that is analyzed due to it being the cryptocurrency with the current largest market capitalization. The studies identified that news should be classified into different themes of news as several studies identify that specific news have significantly more impact as compared to other news sentiments despite having the same sentiment score. News related to regulations and adoption of alternative digital currencies (CBDC) should be prioritized in the classification. The gaps identified include utilization of finance specific dictionaries, exploring the non-linear relationship between news sentiment against prices and exploring different classifications of cryptocurrencies. Another aspect that was seldom explored is the sentiment decay and exploration of different levels of granularities may identify new insights into the relationship.

## 2.5 Literature Summary

This section focuses on summarizing the information from all the prior sections. The literature review identifies that cryptocurrencies operate on a peer-to-peer-based transaction system secured through cryptographic algorithms and exhibit significant price volatility, necessitating robust forecasting models. Prior research has demonstrated that sentiment analysis, particularly from social media platforms like Twitter, plays a crucial role in predicting cryptocurrency prices. Studies have shown that prices of emerging tokens are more sensitive to market sentiment compared to established or meme tokens.

Different segmentation of tokens, such as established (e.g., Bitcoin, Ethereum), emerging (e.g., Polkadot, Chainlink), and meme tokens (e.g., Shiba Inu, Dogecoin), each exhibit distinct characteristics and market behaviors. Established tokens have high market capitalization and acceptance, emerging tokens are newer with high volatility, and meme tokens have low market capitalization with large price swings based on trends.

Research has also highlighted the importance of news platforms in influencing cryptocurrency prices. Techniques such as filtering news based on relevance and novelty, and splitting sentiment analysis between headlines and article bodies, have been used. Studies have identified non-linear relationships between news sentiment and cryptocurrency returns, suggesting that traditional linear models may not fully capture these dynamics. For example, news with higher entropy transfer values significantly impacts cryptocurrency prices, indicating the importance of certain types of news.

The concept of a "Cryptocurrency Uncertainty Index" has been explored, showing that major events, such as political developments, regulatory changes, technological advancements, and cyberattacks, increase price uncertainty. Key events need to be highlighted in news sentiment analysis, and exploring more nuanced methods and separating headlines from body content could provide better insights.

Cross-validation of existing models and techniques has shown that finance-related news sentiment analysis performs best with specific models like FinBert. Studies have used various sentiment analysis dictionaries and have aggregated data on a day-to-day basis to ensure ease of processing.

Critical analysis of the literature indicates that Bitcoin is the primary focus due to its large market capitalization. Studies suggest classifying news into different themes, as specific types of news can have more significant impacts regardless of sentiment scores. Identified gaps include the need for finance-specific sentiment dictionaries, exploring non-linear relationships, and examining sentiment decay and different levels of data granularity.

Overall, the review highlights the complex and multifaceted relationship between news sentiment and cryptocurrency prices, emphasizing the need for advanced and tailored analytical methods to capture the nuances of this dynamic market.

# 3.0 Methodology

This section covers the proposed methodology for this study and details the different phases that it will encompass. Based on the nature of the study, the proposed methodology is OSEMN which is a robust methodology that is split into 5 primary sections.

Figure 1.0 Proposed Pipeline

A diagram of a company

Description automatically generated

The proposed pipeline, encompasses the 5 Primary Phases of OSEMN

* Obtain: Focuses on the extraction of data from the sources and potential transformations required during said processes.
* Scrub: Overall preprocessing steps taken to prepare the data for exploration and modelling in future steps.
* Explore: This phase consists of common Exploratory Data Analysis steps and highlighting potential insights and patterns within the pre-processed data.
* Modelling: This phase consists of identifying the models to be utilized in the study as well as fine-tuning the parameters to ensure more efficient modelling.
* iNterpret: This phase consists of compiling the outputs from the modelling phase and facilitating the discussion of the results of the study.

By utilizing this methodology and data pipeline, a more streamlined process can be taken during the undertaking of this study, ensuring that each step is achievable to a satisfactory degree. The more descriptive explanation of the phases will be expanded on in the following sections of the chapter.

## 3.1 Obtain Phase

This section covers the data sources that this study will utilize and how the data sources will be accessed. The following is the potential data sources for the study.

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Data | Link | Function |
| CoinDesk | News Articles | <https://www.coindesk.com> | One of the largest news platforms which solely focuses on cryptocurrencies. |
| YahooFinance | Price and Volume Data | <https://finance.yahoo.com/markets/crypto/all/> | Offers free access to their API and historical pricing data for cryptocurrency tokens. |

### 3.1.1 News Data Collection

To access the articles provided by CoinDesk, due to the design of the news website, utilizing libraries such as ‘Scrapy’ and ‘BeautifulSoup’ does not meet the requirements to scrape the articles. The primary tool utilized is selenium which allows for individual sessions to be launched, mimicking user movements. The News Data Collection phase is split into two primary sections, article indexing and article scrapping.

Primary data scrapped from news articles include the Article Body, Header, Sub-Header and publication date. This is to address the different granularities of sentiment and identify which level of data is more influencing on the historical price.

#### 3.1.1.1 Article Indexing

The first step to scrape the data is to set the timeframe that is needed to be scrapped and define the types of news. The scope of the timeframe is set to be between May 1st and July 31st , this is to cover all related articles related to the selected coins below.

|  |  |  |  |
| --- | --- | --- | --- |
| Token Name | Token Code | Classification | Market Capitalization |
| Ethereum | ETH | Established | 350 Billion USD ($) |
| Cardano | ADA | Emerging | 12 Billion USD ($) |
| Shiba Inu | SHIB | Meme | 8 Billion USD ($) |

Tokens to be analyzed should include at least one token from each classification (Established, Emerging and Meme) to allow for more generalized insight. This allows for a wider testing of the classified tokens, identifying if different token types have different impacts from different news sources and news types. The articles are then further subdivided into article types with the following articles being scraped:

|  |  |
| --- | --- |
| Classification | Article Focus |
| Business | Articles under this section often discuss topics such as company earnings, mergers and acquisitions, business strategies, venture capital funding, startups, and the overall financial health of companies within the crypto space. |
| Policy | This includes news on how different countries are approaching regulation, compliance requirements, legal battles, and the broader implications of government actions on the crypto industry. |
| Tech | It includes articles on the latest innovations, software updates, protocol improvements, technical deep dives, and advancements in blockchain technology. |
| Markets | This includes price analysis, market trends, trading volumes, market indices, and insights into the performance of various cryptocurrencies and digital assets. |

After setting the timeframe, tokens and article classifications, the articles are then indexed utilizing selenium to gather all articles that meet the criteria.

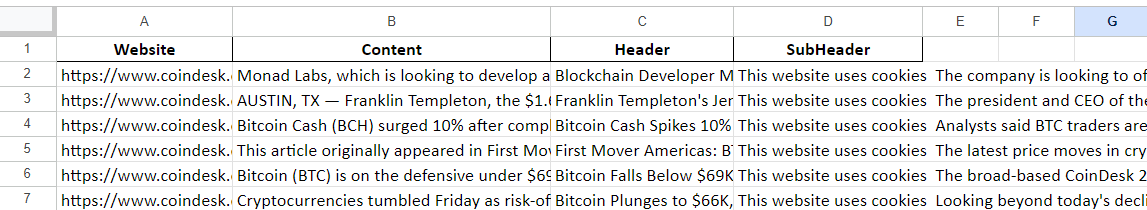
Example of the output of the index scrapper using selenium:



Due to the structure of the website, the scripts utilized were tailored to the specific Cascading-Style Sheet (CSS) tags for scrapping. This data is then stored for the next portion of the study, the article scrapping.

#### 3.1.1.2 Article Scrapping

The article scrapping, also utilizing selenium, runs through the index lists and scrapped the body of the article. Sample output of the article scrapping:



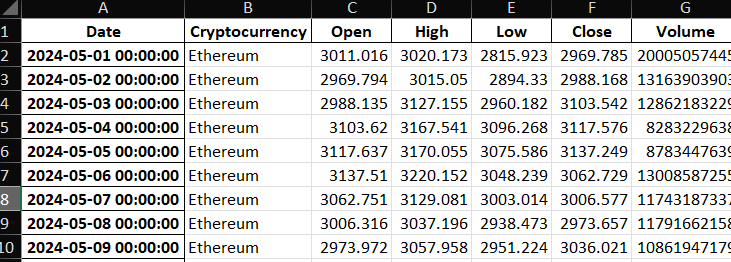
The articles are then segregated by which token they are tagged against (“Bitcoin”,”Ethereum”,”Cardano”,”Shiba Inu”). This data will then be preprocessed and prepared for the generation of the sentiment analysis score.

### 3.1.2 Historical Price and Volume Data Collection

To access the historical prices and volume for Ethereum, Cardano and Shiba Inu, the yahoo finance library “yfinance” is used, due to the cryptocurrency market differing from traditional stock markets has no opening/closing dates, Yahoo Finance uses the same definition of market Open/Closing as CoinMarketCap which defines it as the following:

|  |  |
| --- | --- |
| Data | Description |
| Open | The price of the cryptocurrency at the beginning of a 24-hour period. CoinMarketCap defines the start of day as 12:00AM UTC Time |
| Close | The price of the cryptocurrency at the end of a 24-hour period. CoinMarketCap defines the end of day as 11:59PM UTC Time |
| High | The highest price reached within the period |
| Low | The lowest price reached within the period |
| Volume | The total quantity of cryptocurrency token traded within the period |

After running the script to scrape the historical pricing and volume data, the output is stored into a .xlsx formatted file and a sample of the file is as such:



## 3.2 Scrub Phase

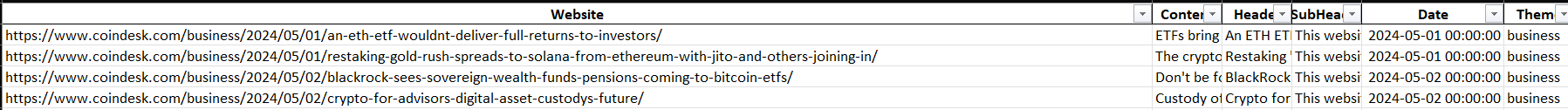
This section focuses on the preprocessing steps that are needed to prepare the data for modelling and exploration.

### 3.2.1 News Data

Once the news data is scrapped, the Date and Theme of the news is needed to be extracted, this data is stored within the URL of the website, the structure of the URL is as follows:

www.<website>.com/<theme>/<year>/<month>/<date>/<article\_title>

From this structure, a regex is used to extract the date and theme and store them as separate columns. The output is structured as follows:



After the date is extracted, the occurrence daily is visualized. This is to have an overall understanding of the distribution of the data and identify any data not within the set timeframe (May 2025 – July 2025) . The visualization is as follows:

A graph with blue lines

Description automatically generated

From the visualization we identify that there are several articles outside the scope, occurring specifically due to hidden links being within the same Div Class as the standard article links. A simple date filter is then done to remove these out-of-scope links, and the visualization is updated as follows:

A graph with blue lines

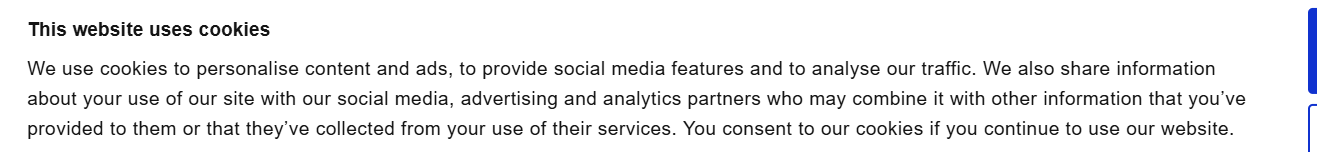
Description automatically generated

Furthermore, to better understand the data, the distribution of articles is then subdivided by the themes of said articles. The visualization is as follows:

A graph with lines and dots

Description automatically generated with medium confidence

Another section of the data that has requires preprocessing is the SubHeader section of the data as due to similar issues, elements that were not required were scrapped such as the notification:



### 3.2.2 Price Historical Data

Due to the nature of cryptocurrencies not having closing dates as compared to traditional stock markets, the pricing data will need to be preprocessed so that a predetermined “closing” time is selected. This is to ensure equal segmentation of the data for further processing and joining with the other datasets.

### 3.2.3 Trade Volume

The data provided by the networks are daily transactions on a token-by-token basis, to generate the trade volume, the data is required to be segmented similarly to the Price Historical Data and aggregated such that all transactions within a given period are grouped daily.

## 3.3 Explore Phase

This section covers the EDA steps to be taken to explore the relationships or patterns within the dataset.

### 3.3.1 Sentiment Analysis Data

The sentiment Analysis data is required to be explored for segmentation of news articles, this is to classify different news and to measure the division and impact of said news. The gaps between news periods need to be identified and filled with sentiment decay scores to further explore the lasting impacts of sentiment.

### 3.3.2 Historical Prices and Trade Volume

Historical Prices will be normalized between the lowest price and highest price, this is to ensure that data is more readable and easier to process.

## 3.4 Modelling Phase

This section covers the modelling phase of the study and the steps and outcomes from the phase.

### 3.4.1 Model Selection

Model selection needs to be conducted separately for the sentiment analysis portion of the study and the price prediction section. Depending on the dictionary used for the sentiment analysis, the outcomes need to be separated and processed in parallel to ensure that the differing insights can be identified (Vader vs FinBert).

## 3.5 iNterpret Phase

This section focuses on the interpretation of the outputs and performances of the models.

### 3.5.1 Sentiment Analysis Modelling

This section focuses on discussing the output of the sentiment analysis section and potential comparison of the dictionaries utilized.

### 3.5.2 Price Prediction

This section focuses on discussing the output of the price prediction model and the model performances.

# 5.0 Results and Discussions

<Placeholder>

## 5.1 Comparative Analysis

<Placeholder>

## 5.2 Hypothesis Validation

<Placeholder>

# 6.0 Deployment

<Placeholder>

# 7.0 Reproducible Study

<Placeholder>

# 8.0 Conclusion

<Placeholder>

# References

Chen, J., Wang, Y., & Liu, X. (2023). The impact of media sentiment on cryptocurrency prices: Evidence from sentiment analysis of news articles. Journal of Financial Markets, 56, 102310. <https://doi.org/10.1016/j.jfinmar.2023.102310>

CoinDesk. (2023). Cryptocurrency news and analysis. Retrieved from <https://www.coindesk.com>

CoinTelegraph. (2023). Blockchain and cryptocurrency news. Retrieved from <https://www.cointelegraph.com>

Liu, Y., & Tsyvinski, A. (2022). Risks and returns of cryptocurrency. Review of Financial Studies, 35(7), 3507-3548. <https://doi.org/10.1093/rfs/hhab093>

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Retrieved from <https://bitcoin.org/bitcoin.pdf>

Vo, A.-D., Nguyen, Q.-P., & Ock, C.-Y. (2019). Sentiment analysis of news for effective cryptocurrency price prediction. International Journal of Knowledge Engineering, 5(2), Article December 2019. <https://www.ijke.org/vol5/116-MK032.pdf>

Bowden, J., & Gemayel, R. (2022). Sentiment and trading decisions in an ambiguous environment: A study on cryptocurrency traders. Journal of International Financial Markets, Institutions and Money, Volume 80, 101622. <https://www.sciencedirect.com/science/article/pii/S1042443122000981?via%3Dihub>

Parekh, R., Patel, N. P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I. E., & Sharma, R. (2022). DL-GuesS: Deep learning and sentiment analysis-based cryptocurrency price prediction. *IEEE Access, 10*, 35398-35409. <https://doi.org/10.1109/ACCESS.2022.3163305>

Koltun, V., & Yamshchikov, I. P. (2023). Pump it: Twitter sentiment analysis for cryptocurrency price prediction. *Risks, 11*(9). <https://doi.org/10.3390/risks11090159>

Rognone, L., Hyde, S., & Zhang, S. S. (2022). News sentiment in the cryptocurrency market: An empirical comparison with Forex. *International Review of Financial Analysis, 69*. <https://doi.org/10.1016/j.irfa.2020.101462>

Banerjee, A. K., Akhtaruzzaman, M., Dionisio, A., Almeida, D., & Sensoy, A. (2022). Nonlinear nexus between cryptocurrency returns and COVID-19 news sentiment. *Journal of Behavioral and Experimental Finance, 36*, 100747. <https://doi.org/10.1016/j.jbef.2022.100747>

Lucey, B. M., Vigne, S. A., Yarovaya, L., & Wang, Y. (2022). The cryptocurrency uncertainty index. *Finance Research Letters, 45*, 102147. <https://doi.org/10.1016/j.frl.2021.102147>

Gadi, M. F. A., & Sicilia, M. A. (2024). A sentiment corpus for the cryptocurrency financial domain: The CryptoLin corpus. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-024-09743-x>

Sapkota, N. (2022). News-based sentiment and bitcoin volatility. *International Review of Financial Analysis, 82*, 102183. <https://doi.org/10.1016/j.irfa.2022.102183>

Rognone, L., Hyde, S., & Zhang, S. S. (2020). News sentiment in the cryptocurrency market: An empirical comparison with Forex. *International Review of Financial Analysis, 69*, 101462. <https://doi.org/10.1016/j.irfa.2020.101462>

Aysan, A. F., Caporin, M., & Cepni, O. (2024). Not all words are equal: Sentiment and jumps in the cryptocurrency market. *Journal of International Financial Markets, Institutions and Money, 91*, 101920. <https://doi.org/10.1016/j.intfin.2023.101920>

Sajter, D., & Kuzmic, M. (2024). Influence of COVID-19 related media and news on the stock and crypto markets. *Ekonomski Pregled, 75*(2), 171–193. <https://doi.org/10.32910/ep.75.2.3>

Helmi, M. H., Çatık, A. N., & Akdeniz, C. (2023). The impact of central bank digital currency news on the stock and cryptocurrency markets: Evidence from the TVP-VAR model. *Research in International Business and Finance, 65*, 101968. <https://doi.org/10.1016/j.ribaf.2023.101968>

Chokor, A., & Alfieri, E. (2021). Long and short-term impacts of regulation in the cryptocurrency market. *The Quarterly Review of Economics and Finance, 81*, 157–173. <https://doi.org/10.1016/j.qref.2021.06.013>

Coulter, K. A. (2022). The impact of news media on Bitcoin prices: Modelling data-driven discourses in the crypto-economy with natural language processing. *Royal Society Open Science, 9*(4), 220276. <https://doi.org/10.1098/rsos.220276>