Machine Learning Models of Crypto Currency Trends Using Attention-based Sentiment Analysis

Abdulla Al Suwaidi

MSc. Thesis

July 2024

A thesis submitted to Khalifa University of Science and Technology in accordance with the requirements of the degree of MSc in Computational Data Science in the Department of Computer Science.

Machine Learning Models of Crypto Currency Trends Using Attention-based Sentiment Analysis

by

Abdulla Al Suwaidi

A thesis submitted in partial fulfilment of the requirements for the degree of

MSc in Computational Data Science

at

Khalifa University

Thesis Committee

Prof. Ibrahim (Abe) M. Elfadel (Main Advisor),
Khalifa University
Dr. Ahmed Altunaiji (RSC Member 1)
Khalifa University
Dr. Andreas Henschel (Co-Advisor),
Khalifa University
Prof. Emilio Porcu (RSC Member 2)
Khalifa University

July 2024

Abstract

Abdulla Al Suwaidi, **"Machine Learning Models of Crypto Currency Trends Using Attention-based Sentiment Analysis"**, M.Sc. Thesis, MSc in Computational Data Science, Department of Computer Science, Khalifa University of Science and Technology, United Arab Emirates, July 2024.

The field of machine learning and the burgeoning world of cryptocurrency share many commonalities, and their integration presents opportunities for valuable insights. Machine learning utilizes algorithms to analyze and predict patterns based on vast datasets. In contrast, cryptocurrency is a digital or virtual currency that leverages encryption techniques for fund transfers and unit generation regulation. By examining the intersection of these two fields, researchers can gain valuable insights into cryptocurrency trends and market behavior. One approach to exploring this intersection involves using attentionbased sentiment analysis in machine learning models. Attention-based models, a type of neural network, identify the most relevant aspects of an input to the output. Sentiment analysis determines and categorises text data's emotional tone, such as news articles or social media posts. Researchers can use these techniques to develop machine learning models that analyze cryptocurrency-related data sources, such as social media or news articles, to identify trends and sentiments. This type of analysis has practical applications for investors, traders, and policymakers interested in cryptocurrency markets.

Sentiment analysis can be used to identify bullish or bearish sentiment among market participants, potentially providing valuable insights into market trends. In contemporary times, machine learning models have emerged as a compelling tool for anticipating future market trends. This is particularly valuable for informed decision-making concerning investment and policy measures. However, using machine learning models to analyze cryptocurrency data presents several noteworthy challenges. The foremost obstacle is the high degree of volatility characterizing the cryptocurrency market. This dynamic makes it difficult to devise precise predictions with machine learning models. Additionally, the decentralized nature of cryptocurrencies and the lack of data standardization

pose considerable challenges to data collection and analysis processes. Despite these challenges, the integration of machine learning and cryptocurrency has the potential to yield valuable insights into market behavior and trends. Attention-based sentiment analysis presents a promising approach to exploring this intersection. Further research in this area can contribute to understanding cryptocurrency markets and their broader financial system implications.

Indexing Terms: Machine Learning, Large Language Models, Natural Language Processing, Crypto Currencies, Crypto Portfolio Management.

Acknowledgement

I would like to express my deepest gratitude to everyone who has supported and guided me throughout the completion of this thesis. First and foremost, I would like to thank my main advisor, Professor Ibrahim Elfadel, at Khalifa University for his unwavering support, encouragement, and insightful guidance. His expertise, patience, and dedication have been invaluable to my research and academic growth. Professor Elfadel's mentorship has provided me with the knowledge and confidence necessary to navigate the complexities of my research topic. His constructive feedback and constant encouragement have played a crucial role in shaping this thesis. I am also immensely grateful to Khalifa University for providing me with the resources and environment necessary to conduct my research. The institution's commitment to academic excellence and research innovation has been a constant source of inspiration. The support from the administrative and technical staff has been instrumental in facilitating my work, and I extend my thanks to everyone who has contributed to this conducive academic environment.
My colleagues at the Abu Dhabi Investment Authority (ADIA) deserve special recognition for their encouragement, feedback, and understanding throughout this journey. Balancing professional and academic responsibilities has been challenging, but the support from my workplace has made it possible. I am grateful to my supervisors and teammates for their patience and for allowing me the flexibility to pursue my academic goals. Additionally, I would like to extend my heartfelt thanks to ADIA for facilitating and providing me with the opportunity to continue my academic journey. Their support and resources have been fundamental in enabling me to achieve this milestone. I would also like to thank my fellow researchers and friends at Khalifa University. Your collaboration, discussions, and shared experiences have enriched my research journey. The camaraderie and intellectual exchange within our group have been incredibly rewarding, and I am fortunate to have had the opportunity to work alongside such talented individuals. To my family and friends, your constant encouragement and support have been my driving force. Your belief in my abilities and your unwavering support have given me the strength to persevere through the challenges. I am especially grateful to

my parents for their endless love and sacrifices, which have laid the foundation for my academic and professional pursuits. This thesis would not have been possible without the collective support of all these individuals and institutions. Thank you for being a part of my academic journey and for contributing to my growth and success. Abdulla AlSuwaidi July 7, 2024

Declaration and Copyright

Declaration

I declare that the work in this thesis was carried out in accordance with the regulations of Khalifa University of Science and Technology. The work is entirely my own except where indicated by special reference in the text. Any views expressed in the thesis are those of the author and in no way represent those of Khalifa University of Science and Technology. No part of the thesis has been presented to any other university for any degree.

Author Name: Abdulla Al Suwaidi

Author Signature: 

Date: July 18, 2024

Copyright ©

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the integration of machine learning and the cryptocurrency market has gained significant attention to gain valuable insights into market behavior and trends. With the cryptocurrency market's meteoric rise and expansion, machine learning has emerged as a powerful tool for analyzing large and complex datasets to identify patterns and trends that may not be readily discernible with traditional analysis techniques. While supervised learning models have been used extensively in the cryptocurrency market to predict future prices based on historical data, natural language processing (NLP) has also gained importance in analyzing large volumes of text data, such as news articles and social media posts, to determine sentiment and extract valuable insights.

In this paper, we explore the potential of attention-based sentiment analysis in machine learning models for predicting trends in the cryptocurrency market. Attention-based models are a type of neural network that identifies the most relevant aspects of an input to the output. Sentiment analysis, a key component of NLP, categorizes text data's emotional tone, such as news articles or social media posts. Using these techniques to develop machine learning models that analyze cryptocurrency-related data sources,

such as social media or news articles, we can identify bullish or bearish sentiment among market participants, providing valuable insights into market trends.

However, using machine learning models to analyze cryptocurrency data presents several significant challenges. The foremost obstacle is the high volatility characterizing the cryptocurrency market, making it challenging to devise precise predictions with machine learning models. Additionally, the decentralized nature of cryptocurrencies and the lack of data standardization pose considerable challenges to data collection and analysis processes.

Despite these challenges, the integration of machine learning and cryptocurrency has the potential to yield valuable insights into market behavior and trends. Attention-based sentiment analysis presents a promising approach to exploring this intersection. This type of analysis has practical applications for investors, traders, and policymakers interested in cryptocurrency markets. We can develop a more comprehensive understanding of the cryptocurrency market and its broader financial system implications using machine learning models and attention-based sentiment analysis.

1.1 Summary about blockchain and cryptocurrency

Blockchain technology and cryptocurrencies have received much attention in recent years due to their potential to revolutionize the financial system. As a distributed and decentralized ledger, the blockchain allows secure and transparent transactions without a central authority, making it a promising technology for various applications, particularly in the financial industry (Nofer et al.; 2017).

Cryptocurrencies, such as Bitcoin and Ethereum, have emerged as digital assets that utilize blockchain technology to enable secure and efficient transactions. Bitcoin, in particular, has been recognized as a new asset class with significant room for expansion, offering attractive investment opportunities (Liu and Tsyvinski; 2021).

However, these technologies also come with significant risks. One of the most significant risks is the possibility of security weaknesses in the blockchain system. Cybercriminals can exploit vulnerabilities in the system, resulting in the loss of funds or sensitive information. Moreover, regulatory unpredictability is another significant challenge as governments struggle to understand and regulate cryptocurrencies, leading to uncertainty about the future of the technology. Market volatility is also a concern, as cryptocurrency prices can be highly volatile, leading to significant losses for investors (Liu and Tsyvinski; 2021).

Fraudulent activity is another risk associated with using blockchain technology and cryptocurrencies. Due to the decentralized nature of blockchain technology, it may be difficult to detect and prevent fraudulent activity. Moreover, the lack of governmental supervision over the use of blockchain technology further complicates the issue (Liu and Tsyvinski; 2021).

The potential applications of blockchain technology and its impact on various sectors appeared in different research. The paper in (Nakamoto; 2008) presents a framework for utilizing blockchain in supply chain management to improve transparency and accountability. On the other hand, (Buterin et al.; 2014) focuses on the economic impact of blockchain technology and argues that it has the potential to disrupt traditional financial systems by providing a more efficient and secure means of conducting transactions. The research (Tasca et al.; 2018) examines the evolution of the Bitcoin economy by extracting and analyzing the network of payment relationships. The study focuses on the

transaction network of Bitcoin and uses a range of network analysis techniques to investigate the structural properties of the network, the flow of funds, and the behavior of Bitcoin users. The authors find that the Bitcoin network exhibits characteristics of both a scale-free network and a small-world network, with a highly centralized core of highly connected nodes. The paper provides valuable insights into the functioning of the Bitcoin economy and highlights the potential for network analysis techniques to shed light on the behavior of digital currencies. Finally, Swan (2015) investigates the effectiveness of using machine learning algorithms for predicting cryptocurrency prices. The study shows that machine learning algorithms can be effective in predicting cryptocurrency prices, but the accuracy of predictions depends on the specific algorithm used. Overall, these research papers provide valuable insights into the potential of blockchain and its applications in different fields.

Despite these challenges, the potential benefits of blockchain technology and cryptocurrencies are significant. As technology advances, it will likely be increasingly used in various applications beyond finance, such as supply chain management, voting systems, and healthcare, among others (Nofer et al.; 2017).

In summary, while blockchain technology and cryptocurrencies have the potential to revolutionize various industries, including finance, they also come with significant risks. To fully realize the potential of these technologies, it is essential to address the challenges and mitigate the associated risks.

1.2 Overview of Machine Learning and NLP

Artificial intelligence (AI) and machine learning (ML) have become increasingly prevalent in the finance industry, allowing for more efficient and effective analysis and decision-making. ML focuses on the development of algorithms that acquire knowledge from data

and use that knowledge to form hypotheses or reach decisions. ML algorithms can analyze large and complex datasets, discovering patterns and connections that may not be readily apparent when using traditional techniques of analytical investigation (Zhou; 2021).

Natural language processing (NLP) is a subfield of ML that focuses on the interaction between computers and human language. NLP algorithms can analyze and comprehend large volumes of text data, such as news articles and social media communications, and draw key insights from the text data that they have processed (Zhang et al.; 2018). Sentiment analysis is an area of NLP that is responsible for determining and extracting feelings and opinions from a body of textual information. In finance, sentiment analysis is often used to evaluate huge quantities of social media posts and news items to determine the mood of the market. Investors may obtain valuable information into future price variations and be able to make better informed investment decisions as a result of monitoring the sentiment of market participants (Medhat et al.; 2014).

Fraud detection is another area where machine learning algorithms have been used. These algorithms can analyze transaction data and identify patterns that are indicative of fraudulent activity (Chen et al.; 2017). In credit scoring, NLP algorithms can analyze credit reports and other financial information to evaluate creditworthiness (Wang et al.; 2018). Portfolio optimization is another area where machine learning can be applied, as algorithms can analyze historical data to identify optimal investment strategies based on risk and return (Du and Tanaka-Ishii; 2020).

In addition to sentiment analysis, machine learning and NLP have a wide range of other applications in finance. Named entity recognition (NER) is another technique used to identify and classify named entities in text, such as people, organizations, locations, and financial instruments. NER can extract information about companies, their

products, and their financial instruments from news articles, social media posts, and other sources of text data (Zhang, Wang, Liu, Zhang and Ji; 2023). Topic modeling is a technique used to identify the topics or themes present in a large corpus of text data. In finance, topic modeling can be used to identify the key topics and trends driving market sentiment and to identify emerging trends in the financial industry (Aziz et al.; 2022). Text summarization is a technique used to automatically generate summaries of large volumes of text data. In finance, text summarization can be used to generate executive summaries of financial reports, earnings calls, and other important documents, which can save analysts time and help them to quickly identify key insights (Abdaljalil and Bouamor; 2021). Machine translation is the process of automatically translating text from one language to another. In finance, machine translation can be used to translate financial reports, news articles, and other sources of text data from different languages, which can help analysts to better understand global financial markets and investment opportunities (Bahja; 2020).

As financial institutions continue to generate vast amounts of data, machine learning and NLP will play an increasingly important role in enabling more efficient and effective analysis and decision-making in the finance industry. With the ability to extract valuable insights from large volumes of data, machine learning and NLP have the potential to revolutionize the way that financial professionals analyze and interpret data, enabling more informed investment decisions and better risk management.

1.3 Background information on sentiment analysis and its use in finance

Opinion mining and sentiment analysis are both terms that refer to the same thing: the study and identification of the emotional tone and viewpoints that are expressed in a piece of text or speech. This method is gaining traction in the realm of financial decision-making as a result of the fact that it has the potential to provide insights into the feelings that investors and the general public have about a certain stock or cryptocurrency.

Attention-based sentiment analysis is a kind of natural language processing (NLP) technique that makes use of transformer models to locate key elements in text data that are essential to analyzing sentiment (Liu et al.; 2018). As contrast to analyzing the whole of the text data, the attention mechanism allows the model to focus its attention just on the sections of the text data that are the most important for determining the underlying sentiment.

In the realm of natural language processing (NLP), a kind of deep learning model known as transformers has been shown to have a great deal of success. "Attention is all you need" by Vaswani et al. revealed the transformer architecture that was developed (Vaswani et al.; 2017). This architecture uses attention processes to analyze input sequences, which enables it to identify detailed data patterns. These patterns may be used to solve problems.

There is a significant amount of untapped potential in the use of attention-based sentiment analysis in machine learning models for the purpose of forecasting market trends in cryptocurrencies. Investors may be able to make more informed decisions about their bitcoin investments if they do attention-based sentiment analysis on massive

quantities of news articles and social media posts, as well as if they monitor large numbers of news articles and social media postings.

There are several uses of attention-based sentiment analysis for the bitcoin market now available. Investors may, for example, identify trends and patterns in market sentiment using sentiment analysis, and then utilize this information to affect the decisions they make about their investments. Moreover, attention-based sentiment analysis may be used to forecast future price movements by making use of historical data, hence enhancing the predictive capacity of machine learning models.

The use of attention-based sentiment research on the bitcoin market has ramifications for the financial investment industry as a whole. Investors may be able to acquire a competitive advantage in the market and make better informed investment decisions by utilizing machine learning algorithms to scan huge volumes of text data and extract crucial insights from that data. When used in machine learning models, attention-based sentiment analysis shows a lot of potential for predicting market trends in the bitcoin business. Investors are able to extract meaningful insights from massive volumes of text data via the use of transformer models and the attention mechanism, and then apply this information to impact their investment decisions. As the cryptocurrency market continues to mature and grow, the significance of machine learning algorithms for investors who are interested in making a profit from this emerging asset class will only grow.

1.4 The purpose of the literature review

This article investigates the use of sentiment analysis as a tool for making investment decisions in both traditional finance and cryptocurrency investing. In particular, we will offer an overview of sentiment analysis in each of these areas, as well as a review of

previous research that has examined how sentiment analysis has been utilized to impact investment decisions.

In the section on traditional finance, we are going to look at the application of sentiment analysis to the study of stock prices, financial news, and data from social media. In this section, we will examine, compare, and contrast the methodology and approaches that were used in these researches, as well as highlight the significant conclusions and restrictions that were discovered. In the next section on investing in cryptocurrencies, we will investigate how sentiment analysis has been used to evaluate the prices of cryptocurrencies, cryptocurrency news, and social media data. In addition, we will evaluate the many distinct approaches and tactics that were used in these studies, as well as discuss the significant findings and limitations associated with them.

The second part of this discussion will focus on the limitations and restrictions imposed by the different approaches to methodology and tactics that are used in sentiment analysis. In addition, we will analyze the similarities and differences between a variety of approaches, methods, and tactics.

In this last section, we will discuss the potential future directions that sentiment research for investment goals might go, as well as the challenges and limitations associated with putting these directions into practice. We will provide an overview of the significant findings as well as recommendations for further research. This literature review comes to a close with some conclusions and recommendations about the use of sentiment analysis in investment decision-making, both in traditional finance and cryptocurrency investing. We are aware of the constraints and challenges that are associated with this topic; yet, we believe that sentiment research has the potential to provide considerable information that may guide investment decisions. Further study is necessary in order to overcome these limitations and investigate potential new directions for using sentiment analysis in

investing decision-making.

Chapter 2

Literature Review

2.1 Sentiment Analysis in Traditional Finance

The term "sentiment analysis" refers to a process used in traditional finance to assess the market's attitude toward a certain company, industry, or the market as a whole. According to the traditional approach to finance, the key sources of sentiment data are articles on the financial market, earnings reports, and analyst reports. In recent years, there has been a rise in the popularity of doing sentiment research on social media websites such as Twitter and StockTwits. Several pieces of research have used sentiment analysis to create predictions about stock prices and guide investing decisions within traditional finance.

A study conducted by (Bollen et al.; 2011) analyzed the sentiment of tweets about the Dow Jones Industrial Average. The researchers discovered that shifts in sentiment effectively predicted the movements that the index would make in the future. Twitter sentiment analysis is used to make stock price predictions, and their predictions were accurate 70% of the time (Mittal and Goel; 2012). In a separate study, (Ren et al.; 2018a; Kurani et al.; 2023) combined sentiment analysis with support vector machine

to successfully predict changes in the stock market. (Ren and Wu; 2018) investigated herd behavior and used sentiment analysis to make predictions on the stock market's direction. Moreover, sentiment analysis has been used in the evaluation of financial news items as well as the projection of stock prices.

According to the findings of (Ding et al.; 2008), a sentiment analysis model that used financial news data performed much better than traditional methods for projecting market returns. (Oliveira et al.; 2016) did sentiment analysis on the data collected from StockTwits and found that it is possible to forecast stock returns based on sentiment and the number of postings. It was discovered by (Bozanta et al.; 2021a) that utilizing transformer models to measure sentiment on StockTwits performed much better than using traditional methodologies for sentiment analysis. Recently, financial sentiment analysis is essential for valuation and investment decision-making, but traditional NLP models often fall short due to their limited parameter size and training datasets. Large Language Models (LLMs), pre-trained on extensive corpora, have shown superior performance in various NLP tasks because of their zero-shot capabilities Zhang, Yang, Zhou, Ali Babar and Liu (2023). However, directly applying LLMs to financial sentiment analysis is challenging due to the mismatch between their pre-training objectives and the task of predicting sentiment labels, as well as the often context-poor nature of financial news. To overcome these issues, we propose a retrieval-augmented LLMs framework, incorporating an instruction-tuned LLMs module for better sentiment prediction and a retrieval-augmentation module for additional context from reliable sources. This approach significantly improves accuracy and F1 score by 15% to 48% compared to traditional models and existing LLMs like ChatGPT and LLaMA. In the realm of traditional finance, sentiment research has shown some promise as a method for predicting stock prices and selecting investments. Nonetheless, the effectiveness of sentiment

analysis may be contingent on the particular context, time range, and data source being assessed. The selected approach and strategy could also change depending on the application and data source being utilized. In addition, there are restrictions placed on sentiment analysis as well as challenges associated with it. One such challenge is that gathering sentiment efficiently and filtering background noise may be challenging.

2.1.1 Overview of sentiment analysis in traditional finance

In conventional finance, sentiment analysis determines how the market feels about a particular firm, industry, or the market as a whole. This may be done for the market as a whole. Articles in financial news publications, earnings reports, and analyst reports are all examples of traditional sources of sentiment data.

2.1.2 Review of studies using sentiment analysis for investment decisions in traditional finance

- **Sentiment Analysis and Stock Prices:**

  Financial analysis is a field that constantly evolves, and traditional methods of forecasting and decision-making may not always be sufficient. This has led to the emergence of new approaches, including the use of sentiment analysis to predict stock market movements. Sentiment analysis involves analyzing textual data to determine the emotional tone of the language used. By applying this technique to social media and other sources of information, analysts can gain insights into the prevailing sentiment of investors and the broader public.

  Several studies have explored the use of sentiment analysis in financial forecasting. For instance, Bollen et al. (2011) demonstrated that sentiment changes in tweets related to the Dow Jones Industrial Average can predict future movements of the

index with up to 87.6% accuracy. Similarly, Zhang (2011) found that incorporating sentiment analysis into a model for estimating daily stock returns of Chinese companies improved the model's prediction accuracy. Other researchers have proposed novel sentiment analysis models for real-time prediction of stock market prices, such as Guo and Li (2019), who introduced a social networks sentiment analysis model based on Twitter sentiment score (TSS). The TSS predicts future market trends up to 15 time samples (30 working hours) in advance, achieving an accuracy of 67.22% for both upward and downward markets. When predicting upward trends, its accuracy under logistic regression and linear discriminant analysis reaches 97.87%. Meanwhile, Lee (2020) used big data from the Daily News Sentiment Index (DNSI) and Google Trends data on coronavirus-related searches to explore the initial impact of COVID-19 sentiment on the US stock market. The study found a correlation between COVID-19 sentiment and 11 select sector indices of the US stock market and suggested strategic investment planning considering time lag perspectives.

In addition to these predictive applications, sentiment analysis has also been used to better understand the factors that drive investment decisions. Parveen et al. (2020) examined the impact of cognitive biases on investment decisions in the Pakistan Stock Exchange (PSX) and found a significant effect of overconfidence and representative heuristic on investment decisions. Overconfidence was found to mediate the relationship between representative heuristic and investment decisions. This study contributes to the growing literature on behavioral finance, which recognizes the impact of cognitive biases on investment decisions. Similarly, Rupande et al. (2019) discussed the importance of incorporating sentiment-driven noise trader activity into asset pricing models to better capture changes in risk

distribution. The article defines noise traders as irrational traders whose preferences for certain stocks are influenced by wants, cognitive errors, and emotions, and whose trading patterns are linked to market-wide sentiment.

Recently, the development and wealth of countries depend heavily on the stock market Agarwal et al. (2023). To analyze stock market data, data mining and artificial intelligence methods are required. One significant factor influencing stock price volatility is the financial success of particular businesses. However, news reports also play a crucial role in determining stock market movements. In this research, we use sentiment classification to leverage non-measurable data, such as financial news articles, to forecast a company's future stock trends. By examining the relationship between news and stock movements, we aim to shed light on the impact of news reports on the stock market. Our study seeks to enhance the understanding of the role of news sentiment in predicting stock market trends.

Overall, the use of sentiment analysis in financial analysis and decision-making is a promising area of research. The technique offers a way to gain insights into investor sentiment and behavior that can complement traditional methods of financial analysis. Moreover, as the studies cited above demonstrate, sentiment analysis can help improve the accuracy of financial forecasting models and better capture the impact of cognitive biases and market-wide sentiment on investment decisions. As such, sentiment analysis is likely to play an increasingly important role in financial analysis and decision-making in the years to come.

- **Sentiment Analysis and Financial News:**

  Sentiment analysis can be used to study various aspects related to financial news, including the identification of the sentiment of the news article (positive, negative, or neutral), the sentiment of the financial instrument being discussed, the sentiment of the company being discussed, and the sentiment of the market as a whole.

It can also be used to study the relationship between sentiment and stock prices and how sentiment affects trading behavior and market movements. Other aspects that can be studied include the impact of different news sources on sentiment and how sentiment changes over time.

Several NLP techniques can be used to address sentiment analysis in financial news. One common technique is to use a bag-of-words model, which involves representing the text as a collection of individual words without considering their order or structure. This model can be augmented with stemming and stop-word removal techniques to improve accuracy further. (Li et al.; 2014) discusses using sentiment analysis to predict stock prices based on financial news articles. The authors argue that previous models have not taken into account the impact of news sentiment on stock price movements. The paper proposes a generic stock price prediction framework incorporating sentiment analysis using two sentiment dictionaries. The models are evaluated using historical data from the Hong Kong Stock Exchange and news articles. The results show that models with sentiment analysis outperform bag-of-words models at individual stock, sector, and index levels and that sentiment polarity is not useful for predictions. Additionally, there is little difference between models using the two sentiment dictionaries.

Another technique is to use machine learning algorithms such as decision trees, random forests, and support vector machines to classify text as positive, negative, or neutral. These algorithms are trained on a labeled dataset of financial news articles and then applied to new articles to predict their sentiment. (Shuhidan et al.; 2018) explores the application of machine learning algorithms in sentiment analysis of financial news headlines in Malaysia, using data from the Business section of the New Straits Times. The study employs the Opinion Lexicon-based algorithm and Naïve Bayes algorithm for sentiment analysis after pre-processing

steps such as stop word removal and stemming to clean the dataset. Sample outcomes of analysis for both algorithms are explained, and the study's conclusion summarizes the findings and suggests future work in the field. The study's results can benefit stakeholders seeking knowledge or data in the financial world.

Deep learning models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have also been used to perform sentiment analysis on financial news. These models are trained on large datasets of labeled text and can capture complex patterns and relationships between words and phrases. (Souma et al.; 2019) investigates the predictive ability of news sentiments based on stock price returns to forecast financial news sentiments. The authors use Wikipedia and Gigaword's five corpus articles from 2014 to create word vectors for inputs into a deep learning TensorFlow network. They apply a combination of deep learning methodologies to train the Thompson Reuters News Archive Data from 2003 to 2012 and test their method on 2013 News Archive data. The authors find that selecting the news with the highest positive and negative scores improves the forecasting accuracy of their methodology. The study shows the potential of using deep learning and sentiment analysis for financial market forecasting. Wu et al. (2022) proposes a new stock price prediction method called S_I_LSTM which incorporates multiple data sources and investor sentiment to improve prediction accuracy. The method involves crawling multiple data sources on the internet, including stock historical data, technical indicators, non-traditional data sources (such as stock posts and financial news), and analyzing sentiment using a convolutional neural network. The results show that the proposed method outperforms traditional methods with a mean absolute error of 2.386835. The paper also discusses the history of stock market prediction theories, the limitations of existing

prediction methods, and the impact of text mining and deep learning on stock prediction. The contributions of the paper include the S_I_LSTM framework, the sentiment analysis method based on CNN, and the demonstration of improved prediction accuracy using multiple data sources and sentiment analysis.

Recent research highlights the importance of sentiment analysis in the financial sector, aiding investors in decision-making Adhikari et al. (2023). Traditional methods struggle with vast data and complex language features, and they often lack interpretability. To address these issues, a novel method using an explainable hybrid word representation was developed, incorporating data augmentation and integrating three types of embeddings to capture linguistic complexities. This representation is processed by a convolutional neural network (CNN) with attention mechanisms, enhancing sentiment detection accuracy. Experimental results show the model significantly outperforms several baselines and offers improved interpretability through visualizations that explain predictions, thus fostering user trust in financial sentiment analysis (Author et al., Year).

Other NLP techniques, such as topic modeling, named entity recognition, and text summarization, can also be used in conjunction with sentiment analysis to gain deeper insights into the underlying trends and drivers of market sentiment. Overall, the use of NLP techniques in sentiment analysis of financial news can provide valuable insights for investors and financial analysts in making informed investment decisions.

In order to provide an accurate prediction of stock prices, financial news stories were subjected to sentiment analysis. According to the findings of (Ding et al.; 2008), a sentiment analysis model that used financial news data performed much better than traditional methods for projecting market returns. Notwithstanding this, a separate study that was carried out by (Tetlock et al.; 2008) found that

the emotional response to financial news did not have any predictive capacity for short-term market gains. The context and length of time being looked at may have a role in determining whether or not sentiment analysis in financial news is effective.

- **Sentiment Analysis and Social Media:**

  In addition to financial news, social media has also become a popular source of information for predicting stock market trends. As pointed out by Li et al. (2019), social media users who demonstrate an interest in finance offer insights into ways irrational behaviors can affect stock markets. Sentiment analysis of Twitter content has been shown to predict trends in stock markets, and computer-based trading systems have been developed to buy or sell based on inferences of mood regarding the market. Incorporating sentiment scores from social media into stock price prediction models has also been explored by researchers. For instance, Huang and Liu (2020) used logistic regression and text mining to incorporate sentiment scores into the model to improve the accuracy of stock price prediction.

  Investor decision-making in the stock market is not solely based on stock-price-related indicators and financial news but can also be influenced by sentiment scores from social media. This finding is supported by Mehta et al. (2021), who analyzed the impact of the representative heuristic and overconfidence cognitive biases on investors' decision making in the Pakistan Stock Exchange. The study found a significant effect of overconfidence and representative heuristic on investment decisions and provided evidence of overconfidence mediating the relationship between representative heuristic and investment decisions. These findings contribute to the literature on behavioral finance and highlight the importance of incorporating investors' emotions and opinions into stock market analysis.

In addition to sentiment analysis, Derakhshan and Beigy (2019) discusses the challenge of opinion mining due to the increasing volume of user reviews on the internet. The authors introduce a part-of-speech graphical model to extract user opinions and test it on two datasets in English and Persian, including the Iranian stock market social network. The model achieves better accuracy in predicting stock market trends compared to methods that use explicit sentiment labels for comments.

Social media generates vast amounts of sentiment-rich data, offering valuable insights into public opinion on various topics Omuya et al. (2023). Traditional sentiment analysis methods, including lexicon and machine learning approaches, face challenges due to noise, high dimensionality, and varying data domains. This study presents a new sentiment analysis model for social media data, incorporating dimensionality reduction and natural language processing with part-of-speech tagging. The model's performance, tested using Naïve Bayes, support vector machine, and K-nearest neighbor algorithms, outperforms two other sentiment analysis models, demonstrating improved effectiveness in sentiment analysis through machine learning techniques.

Overall, these studies demonstrate the growing importance of sentiment analysis in the stock market and how incorporating investors' emotions and opinions can improve the accuracy and efficiency of trading. While classic financial theory assumes that participants make rational choices, unexplained fluctuations arise from irrational behavior. Behavioral finance models the public's behavior at a micro-level and incorporates theories from cognitive psychology. Investors' moods or opinions are important indicators in behavioral financial theory and can determine behaviors in stock markets. Therefore, it is crucial to consider sentiment analysis and other behavioral finance approaches when analyzing the stock market.

2.1.3 Analysis and comparison of the different approaches and techniques used in these studies

Sentiment analysis is a rapidly growing field of research in the financial industry, with machine learning, natural language processing, and lexicon-based methodologies among the most commonly used approaches. According to (Tumarkin and Whitelaw; 2001), machine learning-based sentiment analysis has the highest accuracy in forecasting stock returns. However, the choice of method and technique may depend on the specific application and data source used.

On the one hand, sentiment analysis is used to predict stock market movements by analyzing the emotional tone of textual data, such as social media posts, to gain insights into investor and public sentiment. Novel models, such as the social network sentiment analysis model based on Twitter sentiment score (TSS), have been proposed to achieve high accuracy in predicting future market trends. In addition, sentiment analysis has also been employed to understand the drivers of investment decisions.

On the other hand, researchers focus on sentiment analysis of financial news, exploring various aspects such as the sentiment of the news article, the financial instrument being discussed, the company being discussed, and the market as a whole. To address this, several NLP techniques, such as machine learning algorithms and deep learning models, have been proposed and trained on a labeled dataset of financial news articles to classify text as positive, negative, or neutral.

Sentiment analysis is crucial for understanding human communication and is widely applied in marketing to analyze opinions from social media, news, customer feedback, and corporate communication Hartmann et al. (2023). While lexicon-based methods relate words to sentiment scores, machine learning methods offer higher accuracy but are

more complex. This study proposes an empirical framework to assess the trade-offs between accuracy and interpretability based on research questions, data characteristics, and analytical resources. A meta-analysis of 272 datasets and 12 million sentiment-labeled texts shows that transfer learning models generally outperform lexicons, classifying over 20% more documents correctly. However, their performance may fall short of popular benchmarks. The study emphasizes considering context variables like sentiment class count and text length for realistic performance expectations and provides a pre-trained sentiment analysis model (SiEBERT) with open-source scripts for ease of use.

Overall, sentiment analysis has a promising future in financial analysis and decision-making. By providing insights into investor sentiment and behavior, it can complement traditional financial analysis methods and improve forecasting models' accuracy. Moreover, it can help better capture the impact of cognitive biases and market-wide sentiment on investment decisions.

2.1.4 Summary of key findings and limitations

While sentiment analysis has demonstrated potential in the financial industry, it is important to acknowledge that the success of sentiment analysis depends on various factors. For instance, the effectiveness of sentiment analysis may vary depending on the specific circumstances and time range of the data being analyzed. Some studies have found that sentiment analysis is more effective during times of market volatility and uncertainty, as emotions may play a greater role in investment decision-making. However, during times of market stability, the impact of emotions on investment decisions may be less significant, making sentiment analysis less effective.

Furthermore, the choice of the data source is also important to consider when conducting sentiment analysis. The quality and quantity of data can significantly impact the

accuracy of sentiment analysis models. For instance, social media platforms like Twitter may be a valuable source of data for sentiment analysis due to the volume and speed of information available, but they may also contain a significant amount of noise and irrelevant information. In contrast, news articles from reputable sources may provide more accurate and reliable information but may also be limited in terms of the amount of data available.

Additionally, while sentiment analysis can be a powerful financial analysis tool, it has its limitations and challenges. One of the major challenges is effectively collecting and categorizing emotions, particularly in situations where multiple emotions may be present in a single piece of text. This can be particularly challenging when dealing with subtle emotions or sarcasm, which may be difficult to detect using current sentiment analysis techniques. Another challenge is the issue of context, as the meaning of a word or phrase may change depending on the broader context in which it is used. Finally, the issue of bias in sentiment analysis models is also a concern, as the models may be trained on datasets that do not accurately represent the broader population, leading to inaccurate predictions and conclusions.

In summary, sentiment analysis has shown promise as a tool for financial analysis, but its effectiveness is contingent on a range of factors, including the circumstances, time range, and data source being investigated. It is important to carefully consider these factors when using sentiment analysis in financial decision-making and to be aware of the limitations and challenges associated with this approach.

2.2 Sentiment Analysis in Crypto Investing

The field of finance has extensively used sentiment research to gather insights into the attitude of investors and the general public toward certain assets, including cryptocurrency. In the past few years, the application of sentiment analysis in the cryptocurrency market has been expanding at a rapid rate. As a result, many studies have been carried out to investigate the influence of sentiment on cryptocurrency prices and to forecast the movement of prices in the future.

Research like this was carried out by (Zhang et al.; 2018; Abraham et al.; 2018), who examined the sentiment of tweets connected to Bitcoin and discovered that sentiment had a considerable influence on Bitcoin's price fluctuations. In a research conducted in the same vein, (Alvarez et al.; 2015) employed sentiment analysis to investigate the connection between how people felt about the news and changes in the price of Bitcoin. In addition to news articles and other forms of social media, sentiment analysis has also been used to assess the behavior of the stock market by utilizing the sentiment of tweets and the number of tweets sent. (Oliveira et al.; 2013)discovered that sentiment analysis of StockTwits might be used to anticipate the behavior of the stock market. The authors of the study came to this conclusion after doing research. Similarly, (Mittal and Goel; 2012) conducted research using Twitter sentiment analysis to forecast stock prices.

Methods based on machine learning have also been used in the process of analyzing the sentiment included within Bitcoin-related tweets. Using a technique based on machine learning, (Serrano-Cinca et al.; 2015) analyzed sentiment in Twitter data connected to Bitcoin. They discovered that sentiment analysis might forecast Bitcoin values with a reasonable degree of accuracy. While (Ren et al.; 2018b) used sentiment analysis and support vector machines to forecast the direction of stock market movement, (Ren and

Wu; 2018) used an innovative sentiment analysis approach to measure herd behavior in the stock market.

There is a possibility that the efficacy of sentiment analysis is contingent on the particular situation, coin, and time horizon that are being evaluated. For example, (Dodevska et al.; 2019; Karalevicius et al.; 2018) investigated the usefulness of sentiment analysis in finance. They discovered that the performance of sentiment analysis models differed depending on the kind of data being analyzed, the method used, and the market conditions.

In recent years, cryptocurrency has gained popularity as an investment due to its transparency, independence, and non-transactional nature Chahooki and KJ (2023). Analysts and researchers frequently discuss cryptocurrency futures on social media, influencing investment decisions. This paper presents a framework to help traders understand the opinions of key figures and organizations in the cryptocurrency field. Over six months, sentiments from over 90 prominent Twitter users were analyzed using the Vader open-source tool. The study introduces a user importance factor to weigh tweets based on retweets and comments, rather than follower count, reflecting the current relevance of their opinions. This approach assigns lower importance to opinions that lose relevance over time. Findings indicate that both short-term and long-term market trends in cryptocurrencies can be significantly influenced and predicted by user opinions.

To sum up, sentiment research has shown that it may be useful when making judgments about investments in the bitcoin market. Nevertheless, the efficacy of sentiment research may be contingent on various circumstances, and sentiment analysis is not without its limits and difficulties. For example, the unpredictability and volatility of the bitcoin market is one such constraint. When making judgments about investments in the bitcoin market, it is essential to employ sentiment research in combination with

other analytical tools and methodologies.

2.2.1 Overview of sentiment analysis in crypto investing

Sentiment analysis has emerged as a crucial technique in the world of cryptocurrency investing, providing investors with valuable insights into market sentiment and trends. The explosive growth of cryptocurrencies, particularly in recent years, has led to an exponential increase in the volume of data available for analysis. As a result, sentiment analysis has become an increasingly important tool for investors seeking to make informed decisions based on market sentiment.

One common application of sentiment analysis in cryptocurrency investing is to gauge the general public's and investors' feelings towards a particular coin or token. By analyzing social media and news articles, sentiment research can provide valuable insights into the overall sentiment surrounding a coin or token. Positive sentiment can signal potential growth opportunities, while negative sentiment may indicate a decline in the coin's value.

The process of sentiment analysis in cryptocurrency investing is similar to that in conventional finance. News stories, social media, and other online sources are analyzed for sentiment data. However, due to the decentralized nature of cryptocurrencies, sentiment analysis in this area can be more complex than traditional finance. In addition to analyzing data from conventional sources, such as news articles and social media, investors may also need to consider sentiment data from crypto-specific forums, chat rooms, and messaging apps.

Despite these challenges, sentiment analysis remains a crucial tool for investors seeking to make informed decisions in the fast-paced world of cryptocurrency investing. As the crypto market continues to evolve, sentiment analysis will likely become even more

important in helping investors identify potential risks and opportunities in this dynamic and rapidly growing market.

2.2.2 Review of studies using sentiment analysis for investment decisions in crypto investing

Numerous research has investigated the use of sentiment analysis to forecast the price of cryptocurrencies. This helps to guide investment choices in the cryptocurrency market. For instance, (Abraham et al.; 2018) conducted research in which they examined the sentiment of tweets connected to Bitcoin and discovered that sentiment had a substantial influence on Bitcoin's price fluctuations. The tone of items in the news linked to cryptocurrency has been investigated in several types of research. For example, (Alvarez et al.; 2015) employed sentiment analysis to research the connection between news sentiment changes and Bitcoin price shifts.

- **Sentiment Analysis and Crypto Prices:**

  Sentiment analysis has been increasingly used in the field of crypto investment to predict cryptocurrency prices. Studies have shown that changes in sentiment can provide valuable insights into the direction of the cryptocurrency market. For instance, research by Dimitrios Koutmos (Koutmos; 2018) found that sentiment analysis can be a powerful tool in predicting cryptocurrency prices, particularly in the case of Bitcoin. Another study by Bouri et al. (2017) indicated that sentiment analysis can be used to identify and predict trends in the cryptocurrency market. The research found that sentiment analysis of Twitter data was effective in identifying positive and negative sentiment associated with cryptocurrencies and that this information could be used to predict future cryptocurrency price movements. Additionally, a study by Guo et al. (2021) highlighted that combining multiple

sentiment analysis techniques, including deep learning models, can significantly improve the accuracy of cryptocurrency price prediction models. These studies suggest that sentiment analysis can play a vital role in understanding and predicting cryptocurrency prices, making it a valuable tool for investors and traders.

Predicting Bitcoin price trends is essential as they reflect the broader cryptocurrency market dynamics Jung et al. (2023). Given Bitcoin's relatively short market history and significant price volatility, various studies have investigated the factors influencing its price changes. Previous research has attempted to forecast Bitcoin prices using Twitter data, but these studies were limited by the amount of data and short prediction periods (less than two years). In contrast, this study collected data from Reddit and LexisNexis spanning over four years. By incorporating technical and sentiment indicators into the price data and considering the volume of posts, the performance of six machine learning techniques was assessed. The extreme gradient boosting (XGBoost) model achieved an accuracy of 90.57% and an area under the receiver operating characteristic curve (AUC) of 97.48%. The study demonstrated that combining sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner (VADER) with 11 technical indicators—including moving averages, relative strength index (RSI), and stochastic oscillators—can significantly enhance Bitcoin price trend predictions. Consequently, the input features outlined in this research can be effectively applied to Bitcoin price prediction, enabling investors to make more informed decisions regarding Bitcoin investments.

Furthermore, the growing popularity of sentiment analysis in the crypto investment industry has led to the development of specialized sentiment analysis tools for cryptocurrencies. For example, CryptoMood[1] is a sentiment analysis tool specifically designed for the cryptocurrency market that analyzes news articles,

[1]https://betalist.com/startups/cryptomood

social media posts, and other sources of data to provide sentiment analysis for cryptocurrencies. These specialized tools are likely to become increasingly prevalent as sentiment analysis continues to gain popularity in the crypto investment industry.

Lastly, sentiment analysis has demonstrated its potential as a valuable tool for predicting cryptocurrency prices. The research findings suggest that sentiment analysis can help investors and traders make more informed decisions by providing insights into market sentiment and trends. As the cryptocurrency market continues to evolve, sentiment analysis is likely to become an increasingly important tool in understanding and predicting market movements.

- **Sentiment Analysis and Crypto News:**

  The rise of cryptocurrencies has prompted researchers to explore various techniques to analyze and predict their price fluctuations. One such technique is sentiment analysis, which involves studying news items and social media feeds to gauge the sentiment around a particular cryptocurrency. The potential of sentiment analysis for predicting price fluctuations has been highlighted in various research studies. For instance, Vakil (2019) found that sentiment analysis of cryptocurrency news stories might accurately forecast price fluctuations. However, the bitcoin market's unpredictable nature can make it challenging to do sentiment research. Recent research has proposed new methods for predicting the direction of Bitcoin prices using sentiment analysis. Gurrib and Kamalov (2022) has proposed a novel approach that uses linear discriminant analysis (LDA) and sentiment analysis to predict the direction of BTC prices. The authors train an LDA-based classifier that uses BTC price information and news announcement headlines to forecast the next-day demand for BTC prices. Including news sentiment resulted

in the highest forecast accuracy, suggesting that BTC news sentiment and asset-specific factors are essential in predicting tomorrow's price direction.

In another study, Vo et al. (2019) proposed a two-stage system that uses natural language processing algorithms to analyze news data and produce a sentiment score. The sentiment score is then fed into an LSTM network to predict price movement. The authors argue that sentiment analysis is essential for predicting cryptocurrency prices due to the interactive nature of financial activities. The proposed model improves the accuracy and usefulness of news sentiment data for predicting cryptocurrency prices.

Similarly, Inamdar et al. (2019) examined the impact of social media on the prices of Bitcoin. The authors found that sentiment scores do not significantly impact price prediction unless the scores are not biased towards a particular class. The paper highlights Bitcoin's unique protocol and decentralized nature and the potential for global currency adoption.

Another study by Yao et al. (2019) explores the relationship between news articles and the fluctuation of Bitcoin prices. The authors propose a new text representation method called SentiGraph, which transforms news articles into a graph and achieves superior accuracy in prediction compared to traditional methods. The paper also highlights the importance of news articles in providing information for investors to make informed decisions.

Recently, bitcoin, as the foremost cryptocurrency by market value, plays a crucial role in understanding the broader digital currency marketArjmand et al. (2024). However, due to its historical price volatility, accurately forecasting Bitcoin's value is challenging yet essential. This study focuses on predicting Bitcoin prices using a multifaceted approach encompassing news headline analysis, technical indicators, and historical financial data. Specifically, 3,988 news headlines related to Bitcoin

from the Cointelegraph website between February 7, 2020, and March 8, 2021, were analyzed using CryptoBERT, a transformer pre-trained model tailored for cryptocurrency texts. Additionally, a novel hybrid 2DCNN-GRU deep learning model was developed for price prediction, with parameters optimized using the Taguchi method, a technique based on orthogonal arrays. Comparative analysis against existing deep learning models in the literature underscored the proposed model's superiority, particularly in minimizing Mean Absolute Error (MAE), while achieving competitive results in Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). This integrated approach demonstrates promising advancements in Bitcoin price forecasting, offering potential benefits for investors and researchers alike in navigating the dynamic cryptocurrency market landscape.

To sum up, sentiment analysis has emerged as a valuable technique for predicting cryptocurrency prices. The studies discussed above highlight the potential of sentiment analysis in forecasting the direction of Bitcoin prices. Moreover, sentiment analysis is essential for predicting cryptocurrency prices due to the interactive nature of financial activities. As the cryptocurrency market evolves, sentiment analysis will likely play a more significant role in predicting cryptocurrency prices.

- **Sentiment Analysis and Social Media:**

  The use of social media sentiment analysis in cryptocurrency investment is becoming increasingly popular. Sentiment research on social media lets investors gain insights into the community's feelings towards a specific cryptocurrency. As highlighted by Pano and Kashef (2020), the COVID-19 pandemic has impacted the financial sector, including cryptocurrencies such as Bitcoin. The study examines

the optimal preprocessing strategy for BTC tweets to develop an accurate machine-learning prediction model for Bitcoin prices. It was found that selecting the optimum preprocessing method would prompt machine learning prediction models to achieve better accuracy than actual prices. Raju and Tarif (2020a) aimed to predict the price direction of Bitcoin in USD using machine learning techniques and sentiment analysis on Twitter and Reddit posts. The authors found that LSTM with multi-feature shows more accurate results than the standard method (ARIMA). This highlights the importance of understanding the relationship between public sentiment and Bitcoin price movements to make informed decisions in the cryptocurrency market.

As highlighted by Parekh et al. (2022), the rise of cryptocurrencies and their volatile nature has made it challenging to predict their prices. The authors propose a hybrid and robust framework that considers the interdependency of cryptocurrencies and market sentiments to predict the price of Dash and Bitcoin-Cash using the price history and tweets of Dash, Litecoin, and Bitcoin. The article also mentions the Efficient Market Hypothesis (EMH) and Alternate Market Hypothesis (AMH) and their application in analyzing cryptocurrency market trends and volatility. Several studies, including Sattarov et al. (2020) and Huang et al. (2021), explore the potential of using social media data to develop machine learning models for predicting Bitcoin's price. The use of sentiment analysis in social media to predict the price movement of cryptocurrencies has shown promising results. Prajapati (2020) focuses on using social sentiment as a feature to predict future Bitcoin value, specifically using Google News and Reddit posts. The article found that social sentiment gives a reasonable estimate of how future Bitcoin values may move.

Subbaiah et al. (2024) introduced a novel approach for multimodal sentiment

analysis, addressing the complexities posed by interconnected social media images. Existing methods often analyze individual images independently, overlooking their interrelated nature. They proposed method, the hybrid Arithmetic Optimization Algorithm-Hunger Games Search (AOA-HGS)-optimized Ensemble Multi-scale Residual Attention Network (EMRA-Net), integrates textual, audio, social links, and video modalities to enhance sentiment analysis effectiveness. The AOA-HGS optimizes feature learning, while the EMRA-Net utilizes an Ensemble Attention CNN (EA-CNN) and Three-scale Residual Attention Convolutional Neural Network (TRA-CNN) to capture multimodal sentiments comprehensively. Incorporating Wavelet transform within TRA-CNN mitigates spatial domain image texture loss, and EA-CNN fuses visual, audio, and textual data at the feature level. Evaluation on Multimodal Emotion Lines Dataset (MELD) and EmoryNLP datasets demonstrates superior performance over existing techniques like HALCB, HDF, and MMLatch, achieving higher recall, accuracy, F score, and precision. Moreover, our method proves efficient in computation time across varying training set sizes, highlighting its efficacy for multimodal sentiment analysis in social media contexts.

In summary, sentiment analysis on social media platforms has become valuable for investors to gain insights into community feelings towards cryptocurrencies. The studies discussed in this summary have highlighted the importance of selecting the optimal preprocessing strategy, considering the interdependency of cryptocurrencies and market sentiments, and the correlation between social media sentiment and cryptocurrency price movement. By utilizing sentiment analysis in investment decision-making, investors can gain a reasonable degree of prediction to make informed decisions in the cryptocurrency market.

2.2.3 Analysis and comparison of the different approaches and techniques used in these studies

A wide array of methods and strategies, including machine learning and sentiment lexicons, have been used in the research that has been done on sentiment analysis in crypto investment. For instance, (Serrano-Cinca et al.; 2015) conducted a study in which they analyzed the emotion present in Twitter data about Bitcoin using a method that was based on machine learning. According to the research findings, sentiment analysis has a degree of accuracy comparable to predicting Bitcoin values. On the other hand, the efficacy of sentiment analysis may vary depending on the particular cryptocurrency being examined and the time horizon being taken into consideration.

Another popular approach to sentiment analysis in crypto investment is the use of sentiment lexicons. Sentiment lexicons are dictionaries that contain words and their associated sentiment scores (e.g. positive, negative, or neutral). These lexicons are then used to analyze text and determine its overall sentiment. For example, (Ren et al.; 2018b) developed a sentiment lexicon specifically for the cryptocurrency domain and used it to analyze Bitcoin-related tweets. They found that their lexicon-based approach achieved high accuracy in predicting the sentiment of tweets about Bitcoin. However, some studies have noted that sentiment lexicons may have limited effectiveness in capturing the nuances of sentiment in text, particularly in the fast-paced and ever-changing world of cryptocurrency.

Aside from machine learning and sentiment lexicons, other techniques and strategies have also been used in sentiment analysis for crypto investment. For instance, some studies have used natural language processing (NLP) techniques, such as topic modeling, to extract topics from text and analyze their associated sentiment (Dodevska et al.;

2019). Others have combined sentiment analysis with other analytical methods, such as network analysis, to gain a more comprehensive understanding of the sentiment dynamics in the cryptocurrency market (Vakil; 2019). Overall, while there are many different approaches and techniques used in sentiment analysis for crypto investment, it is important to carefully evaluate their efficacy and limitations to ensure accurate and reliable sentiment predictions.

2.2.4 Summary of key findings and limitations

Sentiment analysis has emerged as a valuable tool for analyzing the emotions and attitudes of investors towards different cryptocurrencies. It has been used in a wide range of studies to help investors make informed decisions, but there are some key findings and limitations that need to be taken into consideration.

One key finding is that sentiment analysis has a degree of accuracy in predicting Bitcoin values and can help investors make more informed decisions about their investments. However, this efficacy can vary depending on the cryptocurrency being examined and the time horizon being considered.

Another limitation of sentiment analysis in crypto investing is the quality and quantity of data used. The accuracy of sentiment analysis is highly dependent on the quality of the data used, and sometimes, the available data may not be sufficient to draw accurate conclusions. Moreover, sentiment analysis cannot account for unexpected events such as major news or regulatory changes that can quickly change the sentiment of investors.

Additionally, the analysis and comparison of the different approaches and techniques used in sentiment analysis show that there is no one-size-fits-all approach to sentiment

analysis. Different techniques and strategies may yield different results and the choice of the technique depends on the specific objectives of the analysis.

In summary, sentiment analysis is a valuable tool for investors in the crypto market, but it is not without its limitations. Careful consideration of the quality and quantity of data used, as well as an understanding of the limitations of sentiment analysis, is crucial to make informed investment decisions.

2.3 Methods and Techniques in Sentiment Analysis

2.3.1 Overview of methods and techniques used in sentiment analysis

Sentiment analysis is a technique used to analyze and classify the sentiment present in a text, which can be useful for various applications, including understanding public opinion about products, services, and events. The use of sentiment analysis has also gained popularity in the field of crypto investing, where sentiment analysis can provide insights into the behavior of investors and the sentiment towards particular cryptocurrencies.

There are various methods and techniques used in sentiment analysis, including rule-based methods, machine learning, and hybrid approaches. Rule-based methods involve creating a set of rules that define the sentiment of a particular text. Machine learning techniques involve training a model on a large dataset of labeled data to predict the sentiment of a given text. Hybrid approaches combine both rule-based and machine learning techniques to achieve higher accuracy.

In addition to the techniques used, sentiment analysis can also involve the use of sentiment lexicons. Sentiment lexicons are dictionaries that contain words and their corresponding sentiment scores. These lexicons can be used to classify the sentiment of a given text by calculating the sentiment score of each word in the text and aggregating the scores.

However, there are limitations to the accuracy and efficacy of sentiment analysis, particularly in the context of crypto investing. The efficacy of sentiment analysis can vary depending on the particular cryptocurrency being examined and the time horizon being taken into consideration. In addition, the quality of the data source used in sentiment analysis can also impact the accuracy of the analysis.

Alslaity and Orji (2024) delved into the critical role of emotion detection and sentiment analysis techniques, pivotal for understanding user polarity and emotions within interactive systems. Emotion recognition not only enhances human-computer interactions by enabling more intuitive interfaces but also facilitates the design of adaptive systems that respond to users' emotional states. With the burgeoning capabilities of machine learning in processing vast datasets, there has been a notable surge in research within this field. To provide a comprehensive overview, they conducted a systematic review encompassing 123 papers on machine learning-based emotion detection. Our analysis reveals several key trends: an increasing interest in this area, predominant use of supervised machine learning algorithms such as SVM and Naïve Bayes, predominant reliance on English language text datasets, and a prevalent use of Accuracy as the primary evaluation metric. These findings highlight the current landscape and inform future directions for developing more responsive and empathetic human-centred systems.

Overall, sentiment analysis can provide useful insights into the sentiment towards cryptocurrencies, but it is important to consider the limitations and potential biases

in the analysis. The choice of method and technique should be based on the specific context and requirements of the analysis.

2.3.2 Discussion of challenges and limitations in sentiment analysis

In spite of the fact that it may be helpful, sentiment analysis has a number of drawbacks and restrictions. For instance, the difficulty of effectively recognizing the sentiment of a text, the prevalence of sarcasm and irony, and the complexity of the text itself may all impact the accuracy of sentiment analysis.

Sentiment analysis in cryptocurrency faces some challenges and limitations. One of the challenges is the cryptocurrency market's volatility (Biju et al.; 2022). The market is highly unpredictable, which makes it difficult to model and predict. Moreover, the lack of regulation and transparency in the cryptocurrency market can lead to misinformation and manipulation, which can affect the accuracy of sentiment analysis results (Zhao et al.; 2021). Another limitation is the language used in cryptocurrency-related discussions, which is often informal and contains jargon and slang (McMillan et al.; 2022). This can make it difficult for sentiment analysis models to classify sentiment accurately. Additionally, sentiment analysis in cryptocurrency may face issues related to data availability, as cryptocurrency-related data is not as widely available as other financial data. The main limitation of Islam et al. (2024) study is the potential bias introduced by the choice of datasets and evaluation metrics. While the proposed deep learning models show promising results across various benchmarks, their performance heavily relies on the specific characteristics and distribution of the datasets used. This can limit the generalizability of the findings to different domains or applications where data characteristics may vary significantly. Finally, the rapid growth and evolution of

the cryptocurrency market can make it difficult for sentiment analysis models to adapt and stay up-to-date (Parekh et al.; 2022).

2.3.3 Comparison of different methods and techniques in sentiment analysis

Many studies have been conducted to evaluate and contrast the effectiveness of various approaches and procedures in sentiment analysis. The findings of these researches may assist investors in selecting the strategy or approach most suited to meet their requirements.

2.4 Future Directions in Sentiment Analysis for Investment Purposes

There is reason to be optimistic about the development of sentiment analysis for use in investing contexts in the years to come. One possible area for expansion is using natural language processing (NLP) and machine learning strategies to achieve greater precision in sentiment analysis. To this end, one possibility is to develop increasingly complex algorithms that can recognize irony, sarcasm, and other types of ambiguous language that can shape attitudes.

Including non-textual data, such as visual material and audio recordings, is an additional path that may be taken in the direction of sentiment analysis. By way of illustration, sentiment analysis may be used in pictures and videos people publish on social media to gather insights into how the general public feels about a particular brand or product. Last, sentiment research might be coupled with other data sources, such as financial and economic statistics, to give a more in-depth knowledge of the elements that

drive investment decision-making. In order to achieve this goal, it may be necessary to construct predictive models that combine sentiment analysis with many other types of data to achieve improved accuracy in investment choices.

Nonetheless, several obstacles and constraints must be overcome to implement these prospective approaches. For instance, the accuracy of sentiment analysis may be impacted by the quality and availability of data and the preconceived notions and personal preferences of the analysts responsible for the analysis. When analyzing non-textual data, such as photos and audio recordings, it is possible that privacy problems and ethical considerations may need to be addressed. This is because these types of data present unique challenges.

2.4.1 Overview of potential future directions in sentiment analysis for investment purposes

The use of sentiment analysis in the field of finance is becoming more popular, and preliminary findings from this line of inquiry have been encouraging concerning their ability to forecast the price movements of various assets (Hung and Alias; 2023). Despite this, there is undoubtedly space for development within the industry. The integration of sentiment analysis with other data sources, such as price and volume data, to improve the accuracy of forecasts is one potential path that may be pursued in the course of future study (Hung and Alias; 2023). This strategy successfully forecasts stock values and might also be extended to predicting prices of other asset types. Additionally, there is a need to explore the potential of deep learning methods in sentiment analysis, which could provide more accurate and nuanced insights into market sentiment (Raju and Tarif; 2020b). This is necessary as there is a need to explore further the potential of deep learning methods in sentiment analysis. In general, sentiment analysis is a fruitful field

of finance study since it offers several future directions that might be further explored and improved.

There are several potential future directions in sentiment analysis for investment purposes. One promising area of research is integrating sentiment analysis with other data types, such as financial and market data (Jin et al.; 2020). Analysts can better understand the underlying factors driving investment decisions by combining sentiment analysis with financial data, such as earnings reports and financial statements.

Another direction is the development of more sophisticated sentiment analysis models that can capture the subtleties of language and context. This could involve using deep learning models like transformer models to learn from large text datasets and identify complex patterns and relationships (Bozanta et al.; 2021b).

Furthermore, there is a growing interest in using sentiment analysis to monitor and predict changes in market sentiment. This could involve the development of real-time sentiment analysis tools that can quickly identify shifts in investor sentiment and alert traders to potential risks and opportunities (Raju and Tarif; 2020a).

Recently, future directions for Alslaity and Orji (2024) research include diversifying language and cultural contexts beyond English, integrating multimodal data sources for richer emotion recognition, enhancing real-time capabilities for interactive systems, addressing privacy and ethical concerns, establishing benchmarks for standard evaluation, exploring unsupervised and semi-supervised learning approaches, and promoting interdisciplinary collaboration to advance understanding and application of emotion detection technologies in human-centered systems. These efforts aim to improve accuracy, applicability, and ethical considerations of emotion recognition systems across diverse domains and user populations.

Finally, sentiment analysis could be used to develop personalized investment recommendations based on individual investor preferences and risk tolerance. By analyzing an investor's past behavior and sentiment, sentiment analysis models could provide tailored investment advice and recommendations that are more aligned with their individual needs and goals.

2.4.2 Discussion of the challenges and limitations in implementing these potential directions

The process of putting the possible possibilities for sentiment analysis into practice meets a number of obstacles and restrictions, the most notable of which are the availability of data and the need for more sophisticated sentiment analysis methods. According to findings from earlier research, the presence of such obstacles might reduce the accuracy with which sentiment analysis can forecast stock prices and other financial events (Poria et al.; 2017). As a result, it is essential to maintain research into novel approaches and strategies in order to circumvent these challenges and enhance the degree of precision achieved by sentiment analysis when applied to the financial app.

2.5 Language Generation Using Large Language Models

Zero-shot and few-shot learning using large language models (LLMs) represent significant advancements in natural language processing. Zero-shot learning enables LLMs to perform tasks without any task-specific training, relying solely on the general knowledge encoded in the model from its extensive pre-training Pourpanah et al. (2022). Few-shot learning, on the other hand, involves training LLMs with a minimal number of examples,

allowing the models to rapidly adapt to new tasks with limited data Wang et al. (2020). These capabilities stem from the broad and diverse datasets used during the pre-training phase of LLMs, which equip them with the ability to generalize across various tasks and domains. The result is a remarkable flexibility and efficiency in applying LLMs to new challenges with minimal additional training, making them highly valuable in diverse applications ranging from text generation to sentiment analysis and beyond.

Using ChatGPT in zero-shot and few-shot learning scenarios Xie et al. (2023) to generate sample text datasets about the positive and negative aspects of cryptocurrencies proves highly beneficial. In zero-shot learning, ChatGPT can generate insightful text on cryptocurrencies without any prior specific examples, leveraging its extensive pre-trained knowledge. This approach allows for the rapid creation of diverse datasets encompassing various viewpoints and topics, such as the benefits of blockchain technology, the potential for high returns, and concerns over security and regulatory challenges. Few-shot learning further enhances this capability by providing a handful of examples, enabling ChatGPT to fine-tune its responses and produce even more contextually accurate and relevant content. This method ensures the generated dataset captures a balanced perspective, aiding researchers, educators, and analysts in understanding and communicating the multifaceted nature of cryptocurrencies. Additionally, the ability to swiftly generate large volumes of high-quality, nuanced text makes ChatGPT an invaluable tool for sentiment analysis, market research, and educational purposes in the rapidly evolving field of cryptocurrency.

In conclusion, the integration of zero-shot and few-shot learning capabilities in large language models like ChatGPT demonstrates the profound potential of these technologies in generating comprehensive and balanced text datasets. These advanced learning techniques facilitate the swift and efficient production of relevant content across various

domains, including the intricate and dynamic field of cryptocurrencies. By harnessing the extensive pre-trained knowledge of LLMs, researchers and practitioners can rapidly develop diverse and insightful datasets, thereby enhancing the depth and breadth of their analyses. This ability to generate high-quality data with minimal input exemplifies the transformative impact of LLMs on natural language processing tasks, promising significant advancements in research, education, and practical applications. As these technologies continue to evolve, their contributions will undoubtedly become increasingly integral to understanding and navigating the complexities of modern data-driven environments.

Chapter 3

Methodology

Sentiment analysis in finance has become increasingly important in recent years. In this study, we aim to investigate the relationship between sentiment and price movements in the context of crypto investing. We first selected benchmark datasets for sentiment analysis and trained several BERT models using transformers and fine-tuning techniques. We then evaluated the performance of the models and chose the best language model for our analysis. Using this model, we performed sentiment prediction on an unlabeled dataset and selected a dataset with corresponding price movements. We matched the sentiment with the prices and drew a figure to visualize the relationship between sentiment and price movements in the crypto market. The results of this study have the potential to provide valuable insights into the dynamics of the crypto market and inform investment decisions. Figure 3.1 summarizes the methodology steps.

3.1 Overview of transformers

Transformers are a type of neural network architecture used for various natural language processing tasks, including sentiment analysis. They have been introduced in the paper

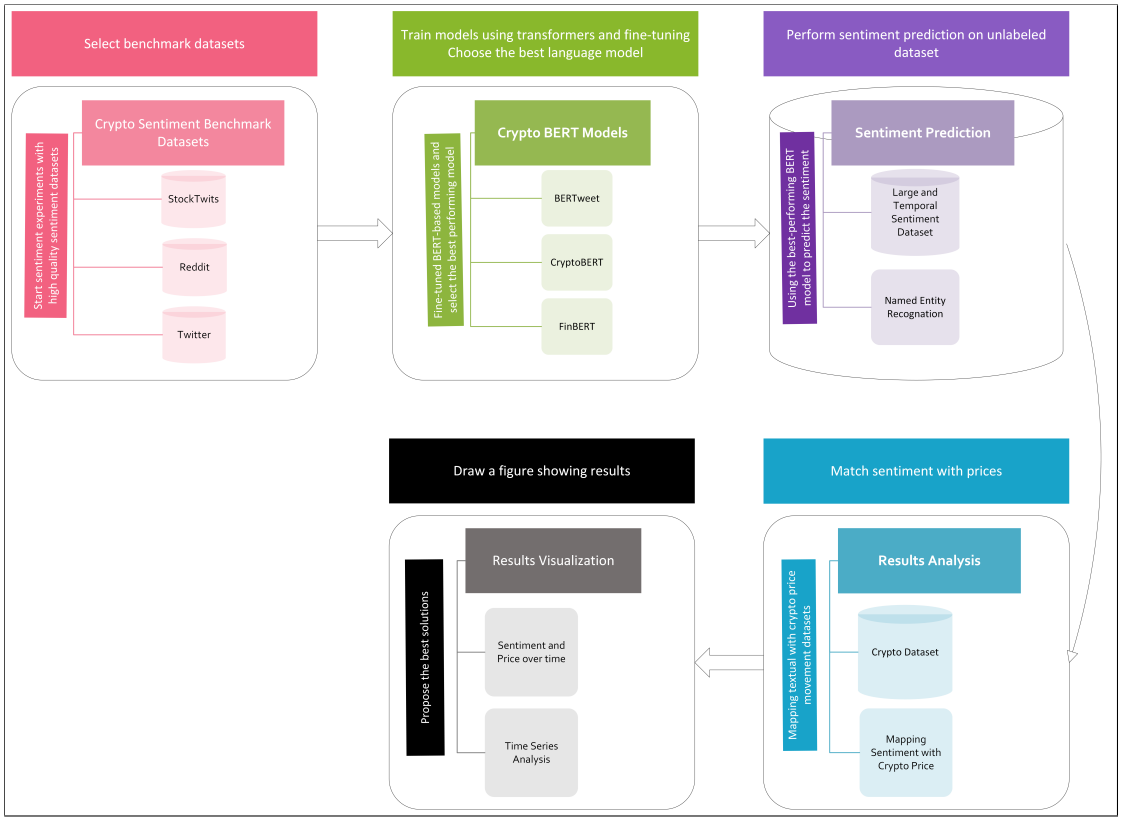


FIGURE 3.1: Crypto Currency Sentiment Analysis System Architecture.

'Attention Is All You Need' by Vaswani et al. (2017) and have since become a popular method for processing sequential data.

Transformers are based on the concept of attention, which allows the model to focus on certain parts of the input sequence that are most relevant to the output. This is achieved through self-attention, where each token in the input sequence is compared to every other token to determine its importance in the sequence context.

Transformers consist of two main components: an encoder and a decoder. The encoder processes the input sequence while the decoder generates the output. Each component consists of multiple layers of self-attention and feedforward neural networks. Figure

3.2 is a superb illustration of Transformer's architecture[1].

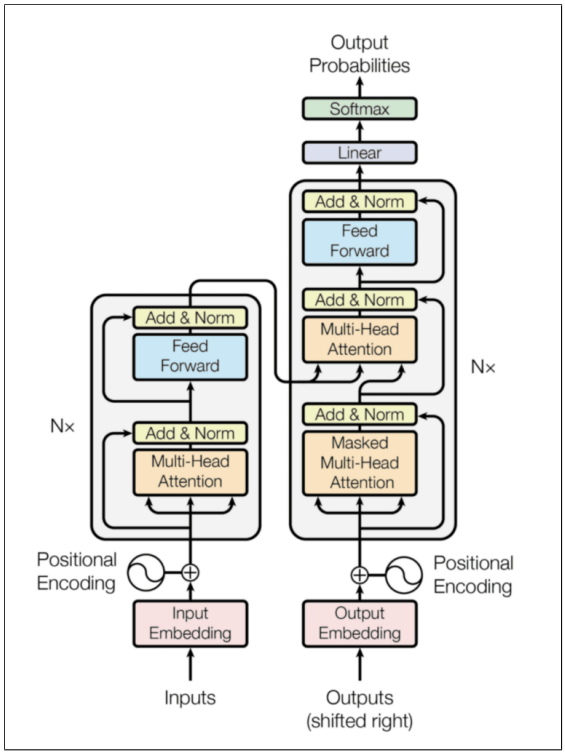


FIGURE 3.2: The Transformer – Model Architecture.

During training, the encoder feeds the input sequence, and the decoder generates the output sequence. The model is optimized to minimize the difference between the generated and desired outputs. Once trained, the model can be used to generate outputs for new input sequences.

In the context of sentiment analysis for cryptocurrencies, transformers can be used to classify the sentiment of text data as positive, negative, or neutral. This is achieved by training the model on a dataset of annotated text data, where each example is labeled with its corresponding sentiment. Once trained, the model can be used to analyze the sentiment of new text data and make predictions about its sentiment.

[1]https://arxiv.org/abs/1706.03762

Overall, transformers are powerful for natural language processing tasks, including sentiment analysis. By utilizing self-attention and multiple layers of neural networks, transformers can capture complex patterns in sequential data, making them ideal for analyzing text data.

3.2 Selection of Crypto Currency Sentiment Datasets

3.2.1 Benchmark Datasets

The first step in our methodology is to select benchmark datasets that are already annotated. This is essential as it helps ensure the model is trained on high-quality and relevant data. The datasets should be diverse in terms of language, sentiment, and domains.

To obtain high-quality training data, it is essential to select benchmark datasets that are already annotated. The selected datasets should be diverse in terms of language, sentiment, and domains to ensure that the models are trained on a wide range of data.

This study conducted a thorough literature review to identify relevant benchmark datasets. Additionally, datasets used in previous studies in sentiment analysis were also considered. The datasets were reviewed and assessed for their quality and relevance, and based on this assessment, the following benchmark datasets were selected:

1. Reddit Crypto Sentiment[2]: The dataset contains 1000 Reddit comments about crypto from Reddit, categorized according to Positive or Negative sentiment.

[2]https://www.surgehq.ai/blog/dataset-of-reddit-crypto-sentiment

2. Bitcoin Tweets 1.4M[3]: Streamed from Twitter API, collecting tweets containing XBT, Bitcoin, BTC.

3. StockTwits-crypto[4]:, Dataset StockTwits-crypto contains all cryptocurrency-related posts from the StockTwits website, from 1st of November 2021 to the 15th of June 2022.

The selected datasets are relevant to the study and provide a diverse range of data in terms of language, sentiment, and domains. These datasets were used to train the BERT models for sentiment analysis, which were fine-tuned using transformers. BERT models are particularly suitable for this task, as they outperform other models in sentiment analysis tasks for natural language processing.

Overall, selecting benchmark datasets is critical in training the sentiment analysis models and ensuring their accuracy and relevance. By choosing high-quality and diverse datasets, the models can be trained on a wide range of data and provide more accurate predictions.

3.2.2 Unlabeled Datasets

Selecting unlabeled datasets from various sources, such as social media, news, or StockTwits, can be an effective strategy to enhance the performance of sentiment analysis models and to perform further analysis on a large time span period. However, manually labeling such large datasets can be time-consuming and expensive. An alternative approach is to leverage the performance of existing models that have performed well on manual or benchmark datasets to automatically label these unlabeled datasets.

[3]https://www.kaggle.com/datasets/paul92s/bitcoin-tweets-14m
[4]https://huggingface.co/datasets/ElKulako/stocktwits-crypto

The automatic annotation process involves using an already-trained sentiment analysis model to predict the sentiment of the unlabeled data. This process is known as transfer learning, where the knowledge learned from one task is transferred to another related task. By transferring the knowledge from an existing model, the sentiment analysis model can learn from the labeled data and enhance its performance on the unlabeled data.

One of the main benefits of using transfer learning for automatic annotation is that it reduces the manual effort required for labeling large datasets. Moreover, by training on diverse unlabeled datasets, the sentiment analysis model can learn to recognize various nuances in sentiment that may be specific to different sources, such as social media or news. This can improve the model's overall accuracy and performance on a wider range of datasets. In summary, leveraging existing models for automatic annotation is a promising approach to enhance the performance of sentiment analysis models on large and diverse datasets.

There are several open datasets available for sentiment analysis of cryptocurrencies in news and social media. One example is to use the news datasets introduced by the common crawl[5]. The datasets can be accessed on AWS S3 in the common crawl bucket at crawl-data/CC-NEWS/. WARC files are released daily and can be identified by the file name prefix which includes the year and month. The datasets are organized as lists of the published WARC files by year and month starting from 2016 until the present. Additionally, authenticated AWS users can obtain listings using the AWS Command Line Interface.

Scrapping, or web scraping, is another method to collect unlabeled text. It has

[5]https://commoncrawl.org/2016/10/news-dataset-available/

emerged as a valuable method for collecting real-time news and data about cryptocurrencies. In the fast-paced world of digital assets, where market dynamics can change in an instant, scraping allows enthusiasts, analysts, and investors to stay well-informed. By automatically extracting information from various online sources, including news websites, forums, and social media platforms, scrapping provides a comprehensive and up-to-date overview of the latest developments in the cryptocurrency space. This method not only facilitates the aggregation of breaking news but also enables the tracking of market sentiments, regulatory changes, and technological advancements. Scrapping plays a crucial role in empowering individuals and organizations to make informed decisions in the dynamic and ever-evolving realm of cryptocurrencies.

3.3 Data Cleaning

The next step after the selection of datasets is to pre-process the text which involves several steps, each of which is necessary for accurate sentiment analysis:

- Data cleaning: Remove unwanted characters, symbols, and stopwords from the text data to reduce dimensionality and improve the efficiency of the sentiment analysis model.
- Tokenization: Break down the text data into individual words or phrases, called tokens, to help the sentiment analysis model understand the context and meaning of each word in the text.
- Conversion to numerical form: Convert the text data into numerical form by using techniques such as vectorization or embedding. This step helps represent each token as a numerical vector or map each token to a high-dimensional space where tokens with similar meanings are located close to each other.

- Feeding into the sentiment analysis model: After the text data is pre-processed, it can be fed into the sentiment analysis model for training and prediction. The trained model can then be used to analyze the sentiment of new text data and make predictions about the sentiment of the text.

- Named entity recognition (NER): it is another crucial step in data cleaning for cryptocurrency sentiment analysis. NER involves identifying cryptocurrency names in the text data. This step is vital since the model needs to differentiate between a cryptocurrency mention and a regular word in the text. Different techniques, such as rule-based and statistical models, can be used to identify cryptocurrency names.

By following these steps, we can ensure that the sentiment analysis model is trained on high-quality and relevant data, improving its accuracy and performance.

3.4 Training and Fine-tuning BERT Models

In this study, we focus on sentiment analysis in the context of crypto investing. We used the second step of training BERT models to achieve this goal, which involves fine-tuning our specific datasets. BERT is a powerful language model capable of understanding the context of words in a sentence, and we utilized the Hugging Face Transformers library to implement the BERT models.

To prepare the data for the models, we tokenized the text and transformed it into a format that can be fed into the BERT model. Fine-tuning the selected benchmark datasets was the next crucial step, as it involved adjusting the pre-trained BERT model to fit the specific sentiment analysis task accurately. This process ensured that the model learned to identify the sentiment of the text with high accuracy.

We trained several models with different hyperparameters to identify the optimal combination of hyperparameters. We evaluated the models' performance using cross-validation, a standard machine-learning model evaluation. Cross-validation involves splitting the data into several subsets, using one subset as a validation set and the rest as the training set. We then assessed the models based on various metrics, such as accuracy, precision, recall, and F1 score, and selected the best-performing model for further analysis.

For our study on sentiment analysis in crypto investing, we will use BERT. This state-of-the-art language model performs superiorly in various natural language processing tasks, including sentiment analysis. We will fine-tune several pre-trained BERT models to better fit our specific sentiment analysis task on crypto-related text data.

The pre-trained BERT models we will use are trained on large amounts of general text data and can be fine-tuned for a wide range of downstream natural language processing tasks, including sentiment analysis. One advantage of using pre-trained models is that they already have a strong understanding of the underlying structure and patterns in natural language, which can reduce the amount of training data required for fine-tuning.

Several pre-trained BERT models have been used in sentiment analysis for finance and investment, including:

1. BERTweet (Pérez et al.; 2021), a pre-trained language model designed explicitly for Twitter data, and FinBERT, a BERT model fine-tuned on financial text data such as news articles and SEC filings. However, since we focus on sentiment analysis in the context of crypto investing, we will select pre-trained BERT models that are more relevant to our domain.

2. FinBERT (Huang et al.; 2022), a BERT model pre-trained on financial communication text. The purpose is to enhance financial NLP research and practice. It is trained on the following three financial communication corpus. The total corpora size is 4.9B tokens.

3. CryptoBERT[6]: a pre-trained NLP model to analyze the language and sentiments of cryptocurrency-related social media posts and messages. It was built by further training the vinai's bertweet-base language model on the cryptocurrency domain, using a corpus of over 3.2M unique cryptocurrency-related social media posts.

Thus, we will consider several pre-trained BERT models, such as BERT-base, BERT-large, and other variations, as well as domain-specific models that have been fine-tuned on crypto-related text data. The optimal pre-trained model(s) selection will be based on their performance on our benchmark datasets, as well as their generalizability and suitability for our specific task. Overall, our focus is to identify the best pre-trained BERT model that can accurately analyze the language and sentiments of crypto-related social media posts and messages.

In summary, we utilized the BERT models to perform sentiment analysis on crypto-related text data. We fine-tuned several pre-trained BERT models with different hyper-parameters, evaluated their performance using cross-validation, and selected the best-performing model for further analysis. We considered several pre-trained BERT models, including domain-specific models, and selected the optimal one(s) based on their performance and suitability for our specific task.

[6]https://huggingface.co/ElKulako/cryptobert

3.5 Fine-Tuning Process LLM Models

The following are the steps for fine-tuning different Large Language Models (LLMs) on a cryptocurrency dataset:

- Model Selection: Choose pre-trained models from Hugging Face's Model Hub based on their architecture and relevance to cryptocurrency data. Models vary in size, performance, and domain focus (e.g., BERT for general NLP tasks, and FinBERT for financial sentiment analysis).
- Dataset Preparation: Tokenize the text inputs from the cryptocurrency dataset to convert them into numerical representations suitable for the model. Format labels or outputs according to the task requirements, such as sentiment labels for classification tasks.
- Model Loading: Load the selected pre-trained models using Hugging Face's Transformers library. Initialize the associated tokenizer to preprocess text inputs during training and inference.
- Task-Specific Configuration: Set task-specific parameters such as batch size (e.g., 32), learning rate (e.g., 10^{-3}), optimizer choice (e.g., Adam optimizer), and number of training epochs (e.g., 20).
- Training Initialization: Initialize the training loop, feeding batches of pre-processed data into the model. Compute predictions, calculate loss against ground truth labels using appropriate loss functions (e.g., cross-entropy for classification), and update model parameters through backpropagation to minimize loss.

- Validation and Metrics Monitoring: Periodically evaluate the model's performance on a validation set using metrics like the F1-score. Implement early stopping techniques if performance on the validation set deteriorates to prevent overfitting.
- Model Saving: Save the fine-tuned models and tokenizers once satisfactory performance metrics are achieved. These saved artifacts can be reused for further experimentation, deployment, or inference on new data.

Thus, several pre-trained LLMs, including domain-specific models fine-tuned on stock-related text data, are considered. The optimal pre-trained model(s) selection will be based on their performance on benchmark datasets, generalizability, and suitability for analyzing cryptocurrency data. Our focus is to identify the best pre-trained LLM that can accurately analyze the language and sentiments of cryptocurrency-related social media posts and news.

In summary, we utilize various LLMs to perform sentiment analysis on cryptocurrency-related text data. We will fine-tune several pre-trained LLMs with different hyperparameters, evaluate their performance using cross-validation, and select the best-performing model for further analysis. We consider several pre-trained models, including domain-specific models, and select the optimal one(s) based on their performance and suitability for our specific task. Figure 3.3 illustrates the fine-tuning process steps.

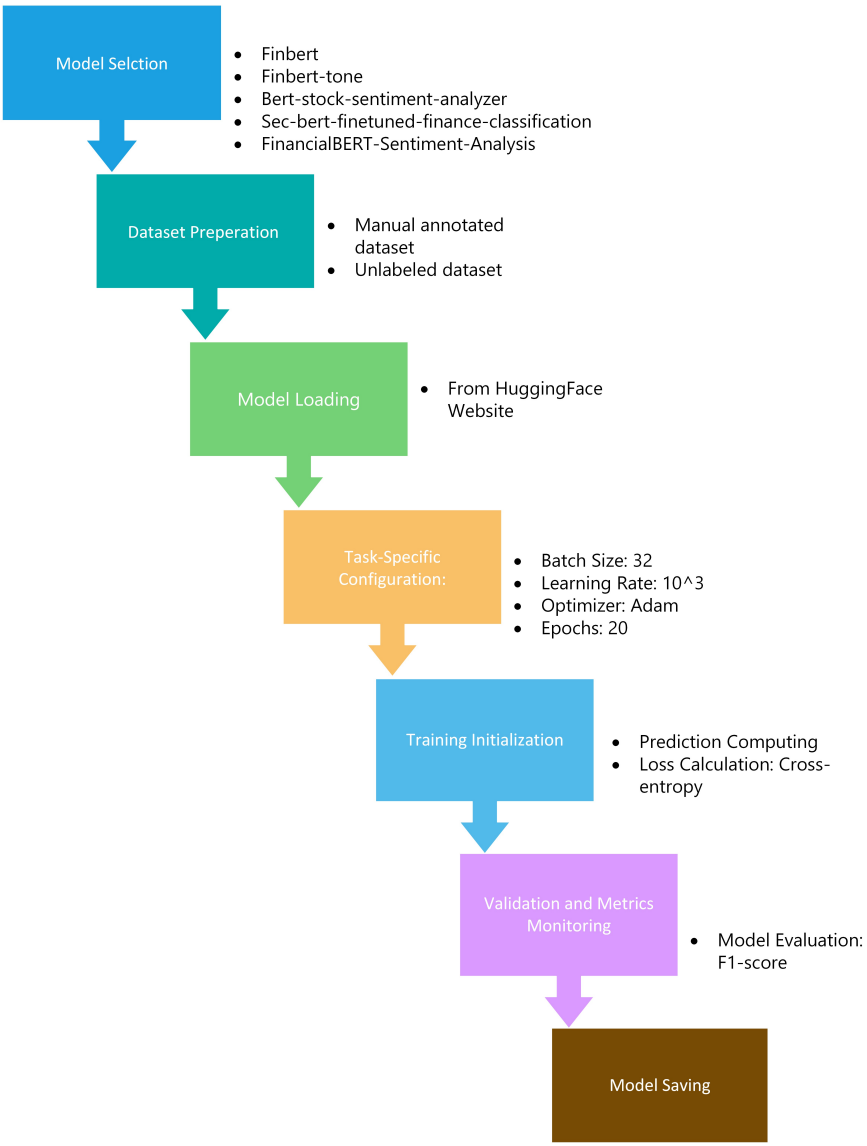


FIGURE 3.3: Fine-tuning process steps.

3.6 Selection of Best BERT Model

To select the best language model, we first trained multiple BERT models using the transformers and fine-tuning method. This method involves fine-tuning a pre-trained BERT model on our benchmark datasets to better fit our specific sentiment analysis

task. We used the Hugging Face Transformers library to implement the BERT models, which is a widely used library for implementing NLP models.

To evaluate the performance of the models, we used several metrics. Accuracy measures the percentage of correctly classified instances, while precision measures the proportion of true positives to all predicted positives. Recall measures the proportion of true positives to all actual positives, while the F1 score is the harmonic mean of precision and recall. By comparing the models' performance on these metrics, we were able to determine the best-performing model for our task of sentiment analysis.

Choosing the best-performing model for our task is essential because it will provide accurate sentiment analysis results. The chosen model will also determine the accuracy of matching sentiment with prices, which is the next step in our analysis.

We also considered the generalizability of the model in choosing the best language model. A model that performs well on the benchmark datasets but does not generalize well to other datasets may not be useful for real-world applications. Therefore, we selected a model that not only performs well on our benchmark datasets but also has good generalizability.

Overall, selecting the best language model is a critical step in our sentiment analysis process. By evaluating the models' performance on various metrics and considering their generalizability, we can choose a model that accurately predicts sentiment and is useful for real-world applications.

3.7 Sentiment Prediction on Unlabeled Crypto Currency Datasets

After selecting the best-performing BERT model for sentiment analysis, the next step is to apply this model to predict the sentiment of an unlabeled dataset. The unlabeled dataset can come from various sources, such as social media, news articles, or user reviews. However, ensuring that the data is relevant to the domain and language of the benchmark datasets used to train the BERT model is important. This ensures that the model can accurately classify the sentiment of the text data.

To perform sentiment analysis on the unlabeled dataset, we first preprocessed the data like the benchmark datasets. This involved tokenizing the text and transforming it into a format that can be fed into the BERT model. We then used the best-performing BERT model to predict the sentiment of each sentence in the dataset.

The model's output is a sentiment score for each sentence in the dataset, representing the text's predicted sentiment. The sentiment score can range from negative to positive, with values closer to zero indicating neutral sentiment. These sentiment scores can be used to analyze the overall sentiment of the text data and identify patterns or trends in sentiment over time.

It is important to note that the accuracy of the sentiment prediction will depend on the quality and relevance of the unlabeled dataset. If the dataset contains text significantly different in domain or language from the benchmark datasets, the model may not predict sentiment well. Additionally, noise or bias in the data may impact the model's accuracy. Therefore, it is important to carefully select and preprocess the unlabeled dataset to ensure the best possible performance of the sentiment analysis

model.

3.8 Selection of Crypto Currency Datasets with Price Movements

Selecting an appropriate dataset containing cryptocurrency price movements is crucial to analyzing the effect of sentiment on crypto prices. The chosen dataset should have accurate and reliable information on the cryptocurrency price movements being analyzed over the same period as the unlabeled dataset.

Several sources of cryptocurrency price data include CoinMarketCap, CryptoCompare, and Yahoo Finance. These sources provide historical price data for various cryptocurrencies, including Bitcoin and Ethereum. We chose CoinMarketCap as our primary source of price data due to its reliability and comprehensive coverage of cryptocurrencies.

Once the price data is collected, it must be preprocessed to ensure consistency with the unlabeled dataset. The preprocessing step involves cleaning the data, removing any missing or erroneous values, and aligning the timestamps with those of the unlabeled dataset.

In addition to the price data, we collected additional data such as market capitalization, trading volume, and social media mentions. This data provides additional context for the analysis and helps identify potential correlations between sentiment and other market factors.

Overall, the selection and preprocessing of the dataset containing cryptocurrency price movements is a critical step in the analysis of the effect of sentiment on crypto

prices. Accurate and reliable data is essential to ensure the analysis's validity and draw meaningful conclusions from the results.

3.9 Matching Crypto Currency Sentiment with Price Movement

Matching sentiment scores with price movements is important in analyzing the relationship between sentiment and cryptocurrency prices. It allows us to see how changes in sentiment impact the price of the cryptocurrency being analyzed. In order to do this, we need to align the time periods of the sentiment and price datasets.

Time series analysis techniques are used to align the time periods of the two datasets. These techniques involve identifying the timestamps in the datasets and ensuring they are consistent across them. Once the timestamps are aligned, the datasets can be merged into a single dataset.

Merging the sentiment and price datasets involves combining the sentiment scores with the cryptocurrency's price movements. This allows us to see how changes in sentiment impact the price of the cryptocurrency over time. We can then use statistical analysis techniques to identify significant relationships between sentiment and price movements.

After merging the datasets, we calculated the average sentiment score for each time period. This allowed us to see how sentiment towards the cryptocurrency changed over time and how those changes corresponded to price movements. We can use this information to gain insights into how sentiment affects the price of the analyzed cryptocurrency.

Figure 3.4 shows the steps of matching the price and sentiment of cryptocurrencies over time.

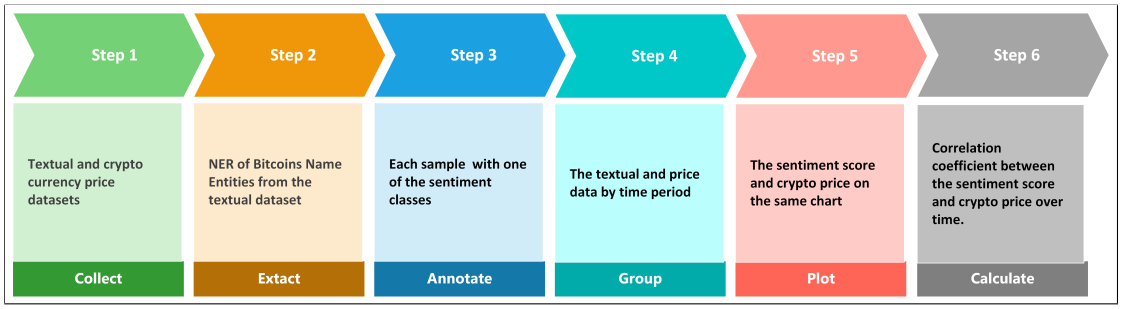


FIGURE 3.4: Steps of matching the sentiment and price of cryptocurrencies over time.

3.10 Visualization of Sentiment and Price Movement Results

When visualizing the relationship between sentiment and cryptocurrency prices over time, we utilized Python's popular data visualization libraries: Matplotlib and Seaborn. These libraries offer various tools to create high-quality charts and graphs, which can help us to understand the relationship between the two variables better.

Matplotlib is a plotting library that allows us to create a wide range of visualizations, such as line charts, scatter plots, and histograms. It provides a high degree of customization for visualizing data, allowing us to adjust colors, fonts, and labels to communicate our findings better. On the other hand, Seaborn is a library built on top of Matplotlib that provides a higher level of abstraction, making it easier to create complex visualizations with fewer lines of code.

We used a line chart to visualize the relationship between sentiment and cryptocurrency prices over time. This allowed us to see how both variables changed over time and

whether any patterns or trends emerged. We plotted sentiment scores on one axis and cryptocurrency prices on the other axis, with time as the independent variable.

We also calculated correlation coefficients to determine the strength of the relationship between the two variables. This allowed us to see how closely related the sentiment scores and cryptocurrency prices were. If the correlation coefficient was close to 1, this indicated a strong positive correlation, meaning that as sentiment increased, so did cryptocurrency prices. Conversely, a correlation coefficient close to -1 would indicate a strong negative correlation; whereas sentiment increased, cryptocurrency prices decreased. There would be no correlation if the correlation coefficient were close to 0, indicating that sentiment scores and cryptocurrency prices were unrelated.

3.11 Datasets

This section describes our main corpus of crypto-currency text and a sub-corpus annotated for sentiment.

3.11.1 Annotated Sub-Corpus

To ensure the accuracy and reliability of our sentiment classifiers, we used the Reddit Crypto Sentiment dataset[7]. The dataset contains 562 Reddit comments about crypto from Reddit, categorized according to Positive or Negative sentiment.

The dataset was annotated by a team of Surgers who are both interested in cryptocurrency and heavy Reddit users. This means that the annotators have a good understanding of the cryptocurrency community and its jargon, which is essential for accurately

[7]https://www.surgehq.ai/blog/dataset-of-reddit-crypto-sentiment

labeling crypto sentiment. The annotators also have experience with data annotation tasks and are familiar with the requirements of high-quality datasets.

The annotators were given a set of cryptocurrency tweets or other social media posts and asked to identify the sentiment of the author (positive or negative). They may also have needed to label other aspects of the tweet or post, such as the topic, the type of language used, or the author's beliefs about cryptocurrency. The annotators' labels were then reviewed by a team of experts to ensure accuracy.

Once the data has been annotated, it can be used to train machine learning models to automatically classify crypto sentiment. These models can then be used to track public sentiment toward cryptocurrency and to identify potential market trends.

In summary, the data annotation of the dataset is likely to be of high quality because the annotators have the necessary skills and experience, and they have a good understanding of the cryptocurrency community and its jargon. Table 3.1 shows sample examples from the manually annotated dataset.

#	**Sentences**
Positive Samples	
1	This is the perfect time to buy. LUNA.
2	Um, what? Bitcoin is constantly being improved and evolving. Taproot was pretty huge.
3	I like the decentralized nature of crypto and the gains are nice too
4	Added 40M on this dip to put me at 380M. Now I only need 120M more to get me to my .5B goal.
5	It scratches my gambling itches and it makes myself learn about investing and economy in general.
Negative Samples	
1	Another ponzi waiting to collapse like Terra
2	The drug dealers and dark web users and the money laundering. They really making crypto look bad.
3	I don't think so crypto recover. All coins going down and down
4	Not gonna trust this horrific guy and never gonna invest on LUNA ever
5	There is no demand tho lol bitcoin has no actual value

TABLE 3.1: Sample positive and negative texts from manually annotated dataset.

3.11.2 ChatGPT

ChatGPT 3.5 and ChatGPT 4 represent significant advancements in natural language processing, with several key distinctions. ChatGPT 4, being the newer iteration, benefits from improvements in both scale and training data. It is a more expansive model, trained on a larger corpus of text, which enhances its understanding of context and language nuances. This results in more contextually relevant responses during interactions. Additionally, ChatGPT 4 exhibits better performance in comprehending complex queries and generating coherent, context-aware responses. The upgraded model is equipped with improved fine-tuning capabilities, making it more adaptable for specific tasks and domains. However, with these advancements comes the challenge of increased computational requirements, which may limit accessibility for some users.

Conversely, ChatGPT 3.5, while not as extensive in scale and training data as ChatGPT 4, still stands as a robust language model. It offers remarkable natural language understanding and generation capabilities, suitable for various applications, from answering questions to content generation. ChatGPT 3.5 is computationally more efficient, making it accessible to a broader range of users. It can efficiently handle a wide array of conversational scenarios and provides valuable insights in both casual interactions and professional applications. Although ChatGPT 4 surpasses it in terms of scale and specific task performance, ChatGPT 3.5 remains a highly valuable tool for those seeking a balance between performance and resource efficiency in their language processing tasks. Ultimately, the choice between the two models depends on the specific needs and resources of the user.

3.11.3 Extension of Annotated Sub-Corpus

To extend the annotated sub-corpus, few-shot learning is used with the help of the ChatGPT to increase the number of samples in the annotated sub-corpus.

We propose a novel approach for extending sentiment annotated sub-corpus using zero-shot and few-shot learning with ChatGPT. ChatGPT is a large language model that can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. We use ChatGPT to generate labeled examples for new sentiment categories. For example, to extend a sentiment dataset to include a new category of "mixed sentiment", we generate examples of mixed sentiment text using ChatGPT.

Leveraging ChatGPT, I extended a manually annotated sentiment corpus through a novel approach involving the identification of positive and negative aspects within the dataset. This innovative method enabled the generation of additional sentences to enrich the sentiment corpus, enhancing its diversity and utility for sentiment analysis research, based on some positive and negative aspects of cryptocurrencies in general and bitcoin in particular. The following is an example of zero-shot prompt used in ChatGPT: Generate 20 positive sentences and 20 negative sentences about the following positive and negative list of cryptocurrencies aspects:

Positive aspects:

1. Decentralization: Not controlled by central authorities.
2. Security: High resistance to fraud.
3. Lower Transaction Costs: Reduced fees.
4. Accessibility: Available to the unbanked.

5. Transparency: Public ledger reduces fraud.
6. Ownership and Control: Direct ownership.
7. Innovation: Versatile blockchain tech.
8. 24/7 Availability: Continuous trading.
9. Financial Inclusion: Provides banking services to the underbanked.
10. Fast Transactions: Swift cross-border payments.

Negative aspects:

1. Price Volatility: High market fluctuations.
2. Regulatory Uncertainty: Changing laws.
3. Lack of Consumer Protections: Irreversible transactions.
4. Energy Consumption: Environmental concerns.
5. Limited Adoption: Not widely accepted.
6. Technological Challenges: Scalability, security.
7. Use in Illegal Activities: Potential for misuse.
8. Market Manipulation: Vulnerable to schemes.
9. Loss of Private Keys: Irrecoverable losses.
10. Competition and Fragmentation: Many cryptocurrencies, leading to confusion.

Few-shot learning played a pivotal role in this process, allowing ChatGPT to comprehend and learn from a limited set of examples. By providing ChatGPT with instances of

both positive and negative sentiment expressions from the existing dataset, the model rapidly adapted to recognize the underlying sentiment polarity and context essential for generating sentiment-related sentences. The following is an example of zero-shot prompt used in ChatGPT: Generate 20 positive sentences and 20 negative sentences in the same context of the following positive and negative samples:

Positive samples:

1. Bitcoin does not need the approval of politicians. Bitcoin is the people's currency supported by the people. Its existence is made legal by the choice of the people. Politicians derive their power from the people and will have to accept the will of the people. That is the way it is.
2. This is pretty big news. Love to see utility even if it's a bearish market overall. Out of fiat to DCA now, might just convert my CCD profits to get some more
3. It's gonna be the currency of the VR world. And we definitely are moving towards that. Giant corporations are focusing their efforts on it (meta, Microsoft). Vr also seems logical from a growth standpoint. From text-only screens to moving pictures, this is the next logical step.
4. This is the perfect time to buy. LUNA.
5. Ever since I read literature about blockchain tech and web3 potential, I am now at 25% crypto exposure. I might or might not be brainwashed by those reads but I assessed my risk and have the conviction that blockchain as tech is here to stay (don't really care about it being a currency, that is not coming anytime soon tbh).
6. Tell them about cryptocurrency so they can invest in bitcoin when they go back
7. No CEO making fake promises = no problems. That's why Bitcoin is king!

8. Interesting. Of course, people who've made money on crypto are going to vehemently disagree with you. I have about 2% of my portfolio in crypto (BTC and ETH). While I fully understand it's risky AF, I strongly believe it will be profitable in the long run.

9. Um, what? Bitcoin is constantly being improved and evolving. Taproot was pretty huge. Preach it brother, glad ltc is finally taking an interest in focusing on privacy too. This is a deciding factor for me, it's why I hold CCD and monero

10. Exactly, remember: the party is just getting started! Join before its too late.

Negative samples:

1. He pretty much scammed billions from the world. I guess Madoff lost to this prick

2. Shitcoins gotta shitcoin

3. I own crypto, and consider it a massive risk. Hell, Tether is probably based on a fiction, and can tank the whole thing. I don't trust any one financial advisor, especially in crypto.

4. Hedgies playing ping pong hoping to pull in retard money along the way

5. Those 100%-useless & stupidly-named greed tokens that boast 1000% APY you see spammed continuously on r/CryptoMoonShots are absolutely not a safe investment.

6. More like 99.9%. Out of 20000 coins thats 20. Im still being very generous. My honest belief is that about 5 coins have actual utility.

7. Wow, glad I cashed out when I did.

8. Crypto itself will implode if Celsius goes down too. ETH will take a beating. Billions will be lost. Now we know why Coinbase made those announcements

9. It became a religion akin to old religions. People defend their coin/token like it was their chosen greek god waging war against other gods. Its honestly very interesting but also very cringe at the same time.

10. The only times I've ever used crypto is for gambling sites and illegal purchases. Even then it was a ball ache to get sorted...

The advantage of employing ChatGPT in this context lies in its inherent ability to understand linguistic nuances and patterns, thanks to its extensive training on a wide array of text data from the internet. This foundation of linguistic knowledge empowers ChatGPT to extrapolate and create meaningful sentences that align with the identified positive and negative aspects.

In practice, the method involved iteratively presenting ChatGPT with positive and negative aspects extracted from the sentiment corpus. These aspects served as guiding examples, enabling the model to generate coherent sentences expressing sentiment in alignment with the identified aspects. This iterative process continued until a substantial number of new sentences were generated, thereby extending the sentiment corpus.

The approach's versatility extends to applications beyond sentiment analysis. Researchers can adapt this methodology to various domains, languages, or specific linguistic contexts by simply providing ChatGPT with appropriate examples of aspects relevant to their research objectives. Moreover, this method is particularly valuable in scenarios where acquiring a large, manually annotated sentiment dataset is resource-intensive or impractical.

In essence, ChatGPT's few-shot learning capabilities, combined with its deep understanding of language, make it a powerful tool for expanding sentiment corpora efficiently. This methodology empowers researchers to diversify and enhance their sentiment datasets, opening doors to more comprehensive sentiment analysis research across languages, domains, and cultures.

We evaluate our approach on a benchmark sentiment dataset, and we show that we can achieve competitive results with significantly fewer labeled examples than traditional machine learning approaches. Our results suggest that few-shot learning with ChatGPT is a promising approach for extending sentiment datasets.

Here are some challenges that we need to address in future work:

- Improving the accuracy of the generated examples.
- Mitigating the biases in the generated examples.
- Developing more efficient methods for generating labeled examples.

We believe that few-shot learning with ChatGPT has the potential to revolutionize the way we develop and maintain sentiment datasets. Table 3.2 shows samples of positive and negative sentences generated by ChatGPT.

3.11.4 Unlabeled Dataset

To perform temporal analysis of crypto prices over time using sentiment analysis, a large and temporal dataset is needed. A large dataset of tweets from Twitter's social media platform is used to perform this task.

#	Sentences
Positive Samples	
1	Cryptocurrencies represent the future of money and finance.
2	Just bought some bitcoin and feeling like a crypto king! #bitcoin #crypto
3	Investing in bitcoin is like investing in the internet in the 90s. Huge potential! #bitcoin #cryptocurrency
4	Bitcoin Hits All-Time High, Surpassing $50,000 Mark
5	NEM's Catapult Upgrade Enhances Speed and Scalability
Negative Samples	
1	Bitcoin facilitates black market transactions and money laundering by enabling anonymous payments.
2	Dump that coin! It's not worth it.
3	Cryptocurrencies are highly volatile and risky investments.
4	The crypto market is filled with scams and fraudulent projects.
5	It's a trap! Don't fall for the Bitcoin hype.

TABLE 3.2: Samples of positive and negative sentences generated by ChatGPT.

3.12 Sentiment Classification

Four different pre-trained language models (finbert-tone, CryptoBERT, FinBERT, and FinancialBERT) were fine-tuned and evaluated in our dataset. The following is a brief description of each pre-trained language model:

- Finbert-tone[8]: FinBERT is a specialized BERT model that has been pre-trained on a substantial 4.9 billion-token corpus of financial communication text, encompassing sources such as Corporate Reports (10-K & 10-Q), Earnings Call Transcripts, and Analyst Reports (Huang et al.; 2023). This pre-training aims to cater specifically to the needs of financial natural language processing (NLP) research and applications. Notably, a model known as "finbert-tone" has been fine-tuned from the FinBERT model using a dataset of 10,000 manually annotated sentences sourced from analyst reports, which are categorized into positive, negative, and neutral sentiments. This fine-tuned model exhibits exceptional performance in

[8]https://huggingface.co/yiyanghkust/finbert-tone

financial tone analysis, making it an invaluable resource for those interested in precisely assessing sentiment within the financial domain.

- CryptoBERT[9]: CryptoBERT is a specialized model derived from the ProsusAI/finbert, a pre-trained NLP model designed for financial text sentiment analysis. This fine-tuned iteration, CryptoBERT, has been adapted for the specific task of sentiment prediction in the cryptocurrency market using the Custom Crypto Market Sentiment dataset. Its evaluation results indicate a loss of 0.3823 on the evaluation set. While CryptoBERT excels in analyzing sentiment within crypto-related content, it's essential to acknowledge that its fine-tuning was conducted on a relatively small corpus of data. During training, specific hyperparameters were employed, including a learning rate of 5e-05, a training batch size of 16, an evaluation batch size of 8, a seed of 42, an optimizer using Adam with betas=(0.9,0.999) and epsilon=1e-08, a linear learning rate scheduler, and a total of 10 training epochs.

- FinBERT[10]: FinBERT is a specialized pre-trained NLP model designed for the analysis of sentiment within financial text (Araci; 2019). It's constructed by extending the BERT language model into the finance domain, a process that involves extensive training with a substantial financial corpus, refining the model specifically for financial sentiment classification. In this fine-tuning process, the model leverages the Financial PhraseBank dataset by Malo et al. (2014). It is essential to consult the paper titled "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models" and a corresponding blog post on Medium for a deeper understanding of its workings. When deployed, the model provides softmax outputs for three sentiment labels: positive, negative, or neutral, facilitating precise sentiment classification in financial contexts.

[9]https://huggingface.co/kk08/CryptoBERT
[10]https://huggingface.co/ProsusAI/finbert

- FinancialBERT[11]: FinancialBERT is a specialized BERT model that has undergone pre-training on an extensive corpus of financial texts[12] (Hazourli; 2022). Its primary objective is to enrich research and practical applications in the financial domain by providing a readily available resource. This model comes fine-tuned for Sentiment Analysis specifically, a task it excels in when applied to the Financial PhraseBank dataset. Comparative experiments have demonstrated its superior performance in contrast to both the general BERT model and other domain-specific financial models. Notably, FinancialBERT was trained on the Financial PhraseBank dataset, which consists of 4,840 Financial News items categorized by sentiment (negative, neutral, positive). During fine-tuning, key hyperparameters like learning_rate, batch_size, max_seq_length, and num_train_epochs were set for optimal results.

[11]https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis
[12]https://huggingface.co/datasets/financial_phrasebank

Chapter 4

Experimental Results

In this section, the experimental results will be presented.

4.1 Dataset Extension

As described in 5.1.2, we used ChatGPT to extend a manually annotated sentiment dataset to improve the classification results. The manually annotated dataset was extended from 562 samples to be 2,492 samples.

4.2 Fine-tuning of Large Language Models

Fine-tuning large language models for sentiment analysis holds paramount importance in enhancing their applicability and performance in understanding and classifying sentiments within textual data. Large language models, such as BERT or GPT, are pre-trained on extensive general datasets, gaining a comprehensive understanding of language structures and nuances. Fine-tuning tailors these models for specific tasks, like sentiment analysis, by exposing them to domain-specific labeled datasets. This process

enables the model to adapt its learned representations to the intricacies of sentiment-related language. The significance lies in the ability to leverage the vast contextual knowledge gained during pre-training while tailoring the model's parameters to the intricacies of sentiment-related tasks. This fine-tuned model, attuned to nuances of positive, negative, or neutral expressions, proves instrumental in various applications, from market sentiment analysis to understanding user feedback, contributing to more accurate and context-aware sentiment predictions.

To perform sentiment classification, both datasets, the manually annotated and the extended dataset, separately. Both datasets were split into training (70%), development (15%), and testing (15%). We use accuracy and both F1-score, macro average, and weighted average as our evaluation metrics. Both datasets were utilized in conducting supervised classification experiments.

BERT Model	**Accuracy**	**F1-Score (macro avg)**	**F1-Score (weighted avg)**
Manually annotated dataset			
Finbert-tone	0.77	0.77	0.77
CryptoBERT	**0.90**	**0.90**	**0.90**
FinBERT	0.71	0.71	0.71
FinancialBERT	0.79	0.79	0.79
Extended dataset			
Finbert-tone	0.89	0.89	0.89
CryptoBERT	**0.93**	**0.93**	**0.93**
FinBERT	0.91	0.91	0.91
FinancialBERT	0.86	0.86	0.86

TABLE 4.1: supervised sentiment classification results using four pre-trained language models.

Table 4.1 shows the classification results of sentiment using four different pre-trained language models. CryptoBERT model outperforms all other models in both, the manually annotated and extended datasets. It achieves 90% and 93% accuracy and F1-Score

in both datasets respectively.

4.3 Best-performing Model Selection

In the process of selecting the optimal fine-tuned model for sentiment analysis, a pivotal stage involves evaluating the models based on their performance on human-annotated samples in the validation set. The fine-tuned models underwent training on a sentiment analysis dataset, and their efficacy was rigorously validated using manually annotated samples. This meticulous validation, which includes key metrics such as accuracy, precision, recall, and F1 score, serves as a test for assessing how well each model generalizes to real-world sentiments. The obtained results showcase the excellence of the best-performing model, Cryptobert, with an accuracy of 75%, precision of 100%, recall of 68.75%, and an F1-score of 81.48%. These results further solidify the selection of Cryptobert as the optimal model, highlighting its remarkable ability to not only accurately classify sentiments but also align closely with human judgment.

4.4 News Scrapping

The process of crawling an unlabeled dataset of news from bitcoin.com and cryptopotato.com involved a comprehensive web scraping strategy. Utilizing web scraping tools or libraries, the crawler systematically navigated through the websites, extracting news titles within the specified timeframe from 2018 to the present. Selenium and Python used to accomplish the crawling process of news. Selenium is a web testing tool that can be used for web scraping by simulating user interactions with a browser. In this case, Python can serve as the scripting language to control Selenium. The process typically involves initiating a browser session, navigating to the desired website, and programmatically interacting with the HTML elements to extract the relevant information. For bitcoin.com,

the HTML elements containing the news articles, titles, and other pertinent details such as date of publish and publisher were extracted. The collected dataset exceeded 40,000 news titles, providing a rich source of information for sentiment analysis. Figure 4.1 shows the distribution of news titles over time in daily, weekly, and monthly basis. The figure also shows that there are an exponential increase in news over time.

The normalized sentiment values were calcualted as follows:

$$\text{Normalized Values(Date)} = \frac{\sum_{i=1}^{n} \text{Value}(\text{Date}_i)}{n}$$

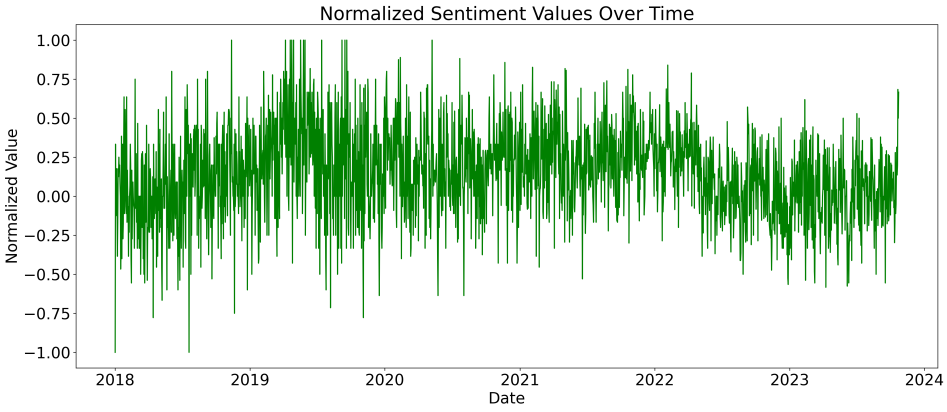
where: **Date** represents the date (daily, weekly, or monthly) in the dataset, **Value(Date_i)** is the sentiment value on a specific date (daily, weekly, or monthly), and **n** is the total number of news on that date (daily, weekly, or monthly).

4.5 Sentiment Prediction

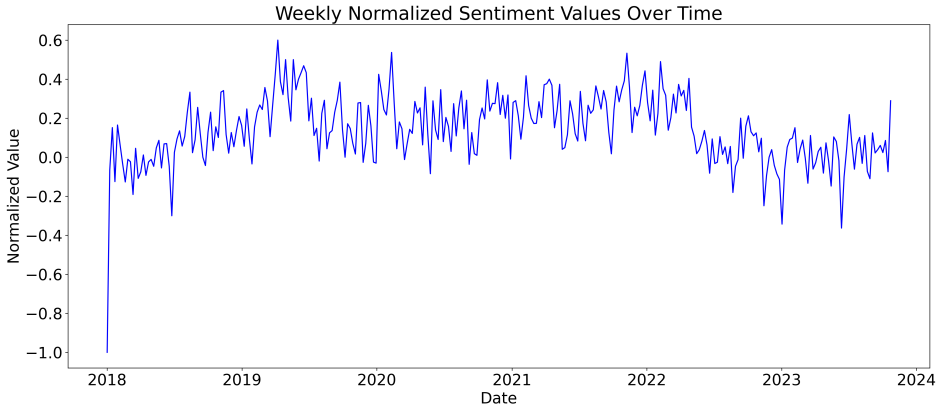
Sentiment annotation of an unlabeled dataset of cryptocurrency news was efficiently performed using the best-performing model, CryptoBERT. Leveraging the strengths of CryptoBERT, a fine-tuned version of the BERT model, on a sentiment analysis task, the model demonstrated exceptional accuracy and nuanced understanding of cryptocurrency-related sentiments. By utilizing transfer learning, CryptoBERT effectively predicted sentiment labels, distinguishing between positive and negative tones within the vast dataset. This sentiment annotation process facilitated a comprehensive analysis of sentiment trends across diverse cryptocurrency news sources, shedding light on market sentiments and contributing valuable insights to stakeholders, investors, and researchers in the cryptocurrency domain.

Subsequently, each news title in the dataset underwent sentiment annotation using the best-performing fine-tuned model, which, in this case, is CryptoBERT. This model, having been fine-tuned specifically for sentiment analysis on cryptocurrency-related text, demonstrated its prowess in discerning positive or negative sentiments within the news corpus. The annotations added valuable labels to the dataset, allowing for a nuanced analysis of sentiment trends in the cryptocurrency domain over the specified time period. This process not only contributed to a deeper understanding of market sentiments but also provided a labeled dataset for further machine-learning applications and insights.

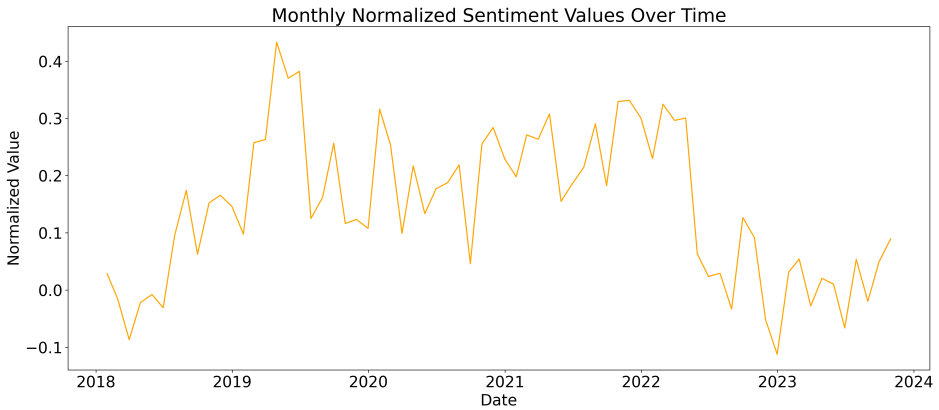
In Figure 4.2, the sentiment distribution is depicted over time, offering insights into the daily, weekly, and monthly trends. This visualization provides a dynamic perspective on the fluctuating sentiments, with discernible patterns in the ebb and flow of positivity and negativity. For instance, a noticeable surge in positive news is evident in the data for April 2019, reflecting a period marked by optimism or favorable developments. Conversely, January 2023 exhibits a spike in negative news, suggesting a phase characterized by challenges or unfavorable occurrences. This comprehensive overview allows for a nuanced understanding of sentiment dynamics across different timeframes, facilitating a more insightful analysis of the evolving narrative.



(A) Daily



(B) Weekly



(C) Monthly

FIGURE 4.2: Normalized Sentiment Distribution Over Time.

4.6 Predictions Analysis

Analyzing the trends in the provided sentiment monthly values over time in Figure 4.2 reveals interesting patterns in the sentiment towards a specific subject, presumably related to cryptocurrencies. The sentiment values, which range between -1 and 1, indicate the overall positivity or negativity of the sentiment during each respective month. Figure 4.3 shows the distribution of positive and negative news over the years.

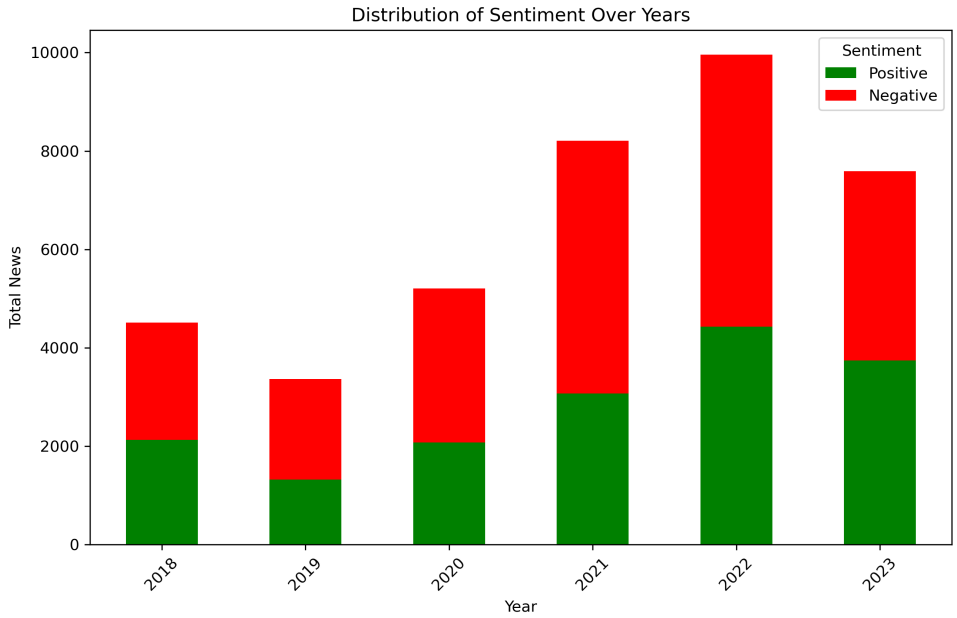


FIGURE 4.3: Distribution of Sentiment of News Over Years.

It's essential to consider external factors, such as market trends, regulatory changes, technological advancements, and major news events, to contextualize these sentiment trends accurately. Analyzing sentiment over time provides valuable insights into the evolving perceptions and emotions surrounding the subject of interest.

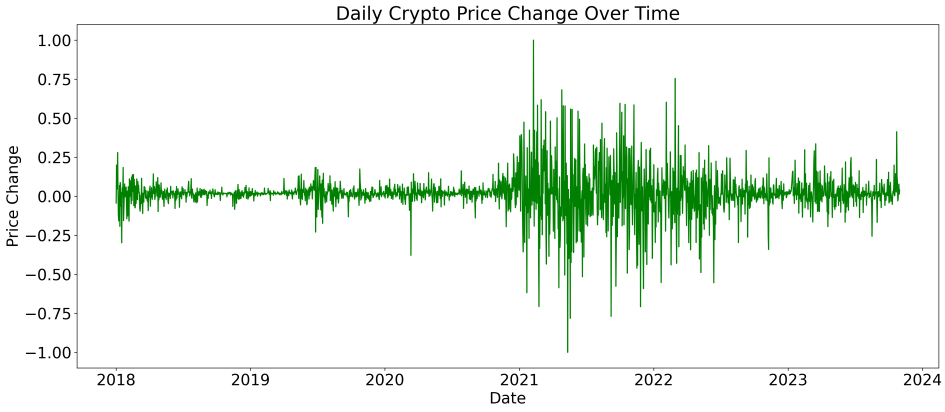
4.7 Crypto-currency Price Datasets

The dataset under consideration encompasses historical Bitcoin and Ethereum price data derived from Yahoo Finance [1]. The price data is collected at a high-frequency level, providing a temporal perspective on market fluctuations. This dataset aims to unravel patterns, trends, and potential relationships between market dynamics and the sentiment expressed in news media, contributing to a comprehensive understanding of factors influencing the cryptocurrency's valuation. Figures 4.6 and 4.5 show the distribution of Bitcoin and Ethereum prices on a daily, weekly, and monthly basis. Price change was normalized to be between 1 and -1 values. The normalized values were calculated as follows:

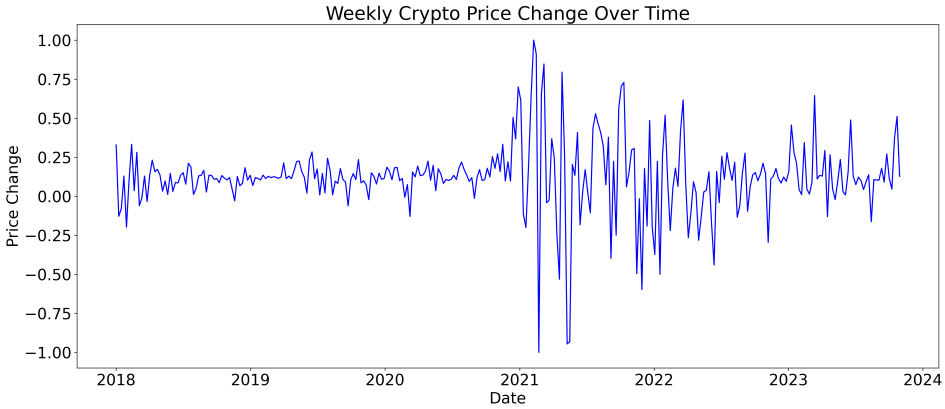
$$\text{normalized\_change} = \left(\frac{\text{Close} - \text{Open}}{\max(\text{Close} - \text{Open}) - \min(\text{Close} - \text{Open})} \right)$$

where: **Open** represents the price at which a cryptocurrency opens at a specific time period, and **Close** represents the price at which a cryptocurrency ends up at a specific time period.

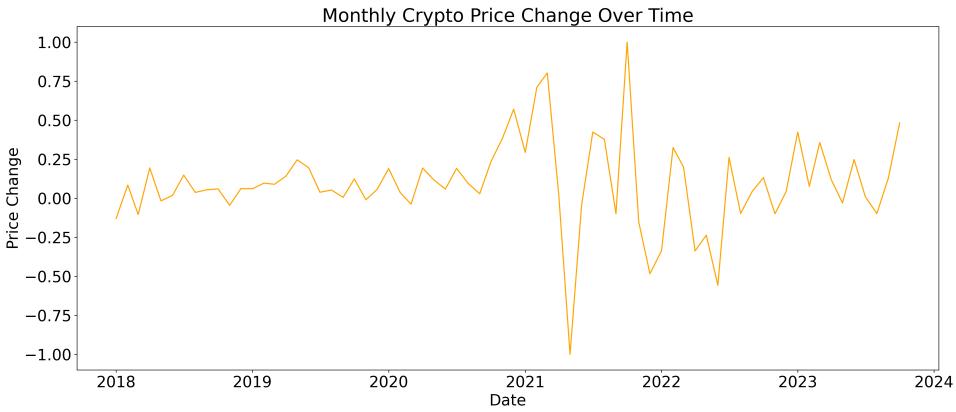
[1] https://finance.yahoo.com/



(A) Daily

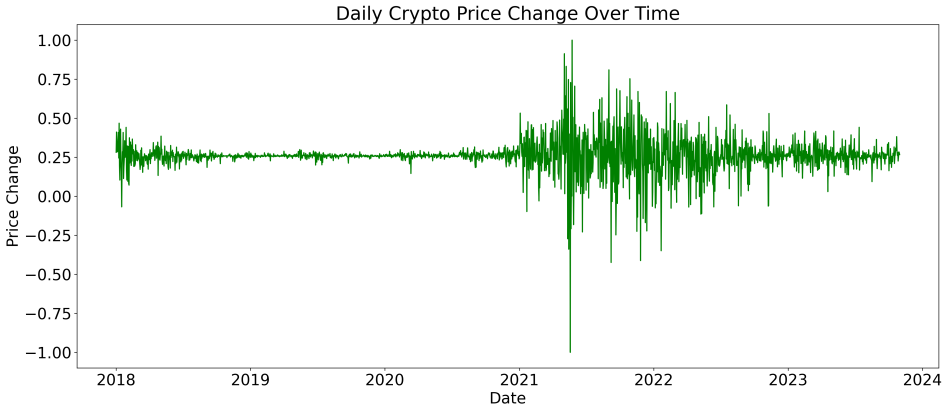


(B) Weekly

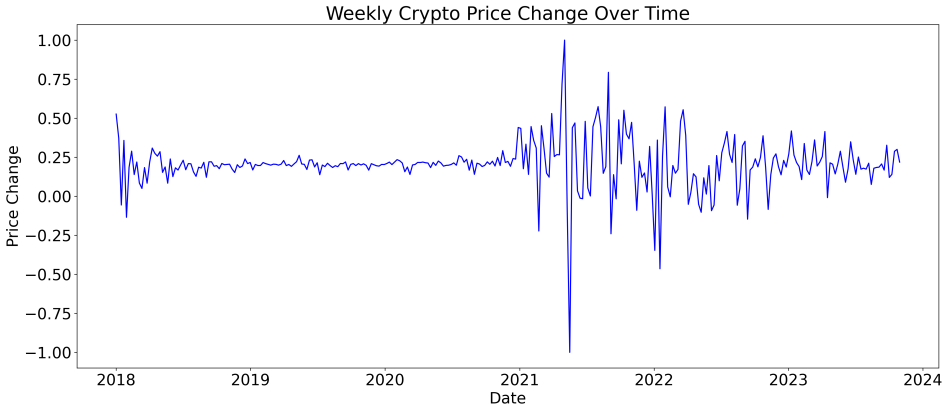


(C) Monthly

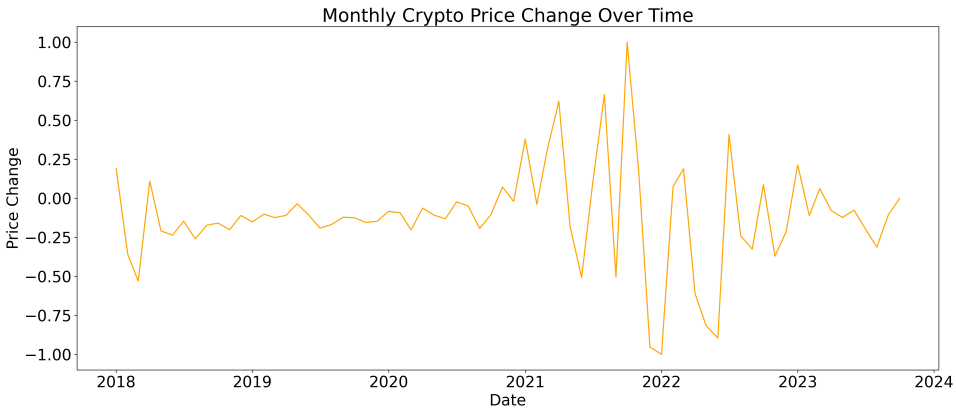
FIGURE 4.4: Normalized Bitcoin Price Distribution Over Time.



(A) Daily



(B) Weekly



(C) Monthly

FIGURE 4.5: Normalized Ethereum Price Distribution Over Time.

4.8 Sentiment and Price Correlation

To compare the sentiment and price of bitcoin, monthly data is used to observe the correlation of both sentiment and price of BTC and ETH over time, as shown in Figure 4.7. Table 4.2 shows the correlation between the crypto price and sentiment of both, Bitcoin and Ethereum.

Crypto Currency	**Correlation**
Bitcoin	0.764
Ethereum	0.734

TABLE 4.2: Correlation between price and sentiment.

4.8.1 Bitcoin

Based on the monthly values of Bitcoin sentiment and price from 2018 to 2023, the following are some potential trends and observations:

Price Trends:

- Volatility Fluctuations: The Bitcoin price shows periods of both volatility and stability, with some months experiencing sharp price fluctuations, while others show relatively stable trends.
- Bullish Periods: There are instances where Bitcoin prices exhibit bullish trends, characterized by consistent positive monthly returns. This may be associated with increased investor confidence, positive market sentiment, or significant events favoring Bitcoin adoption.

- Bearish Corrections: Conversely, there are periods of bearish trends with negative monthly returns. These corrections may be triggered by market uncertainties, regulatory concerns, or profit-taking activities.
- Overall Upward Trajectory: Despite short-term fluctuations, the general trajectory of Bitcoin prices seems to show an overall upward trend over the analyzed period, suggesting a positive long-term sentiment towards the cryptocurrency.

Sentiment Trends:

- Positive Correlation with Price: There appears to be a positive correlation between Bitcoin prices and sentiment. Months with higher sentiment values often coincide with periods of rising Bitcoin prices, indicating a potential relationship between market sentiment and price movements.
- Extreme Sentiment Events: Peaks or troughs in sentiment values may signal periods of extreme market sentiment. These events could coincide with significant price movements, serving as potential indicators of sentiment-driven market shifts.
- Overall Positive Sentiment: The majority of sentiment values seem to lean towards the positive side, suggesting an overall positive sentiment towards Bitcoin over the analyzed period. This aligns with the general upward trajectory observed in Bitcoin prices.
- Stable Sentiment Periods: Some months exhibit relatively stable sentiment values, indicating a consistent sentiment environment. This stability could be influenced by factors such as regulatory clarity, technological advancements, or sustained market confidence.

General Observations:

- Market Sensitivity: Bitcoin prices and sentiment seem to be sensitive to external factors, including regulatory developments, macroeconomic conditions, and technological advancements in the blockchain space.

- Long-Term Investment Perspective: The data suggests that, despite short-term market fluctuations, there is a prevailing positive sentiment towards Bitcoin from a long-term investment perspective.

- Integration of Sentiment Analysis: Integrating sentiment analysis with price trends provides a more holistic understanding of market dynamics. Extreme sentiment events may serve as potential indicators for traders and investors.

4.8.2 Ethereum

The analysis is performed on sentiment and price of Ethereum crypto-currency data from 2018-2023. The following are the observations and analysis:

Price Trends:

- Volatility (2018-2019): Ethereum started the period with significant price volatility, experiencing both positive and negative monthly returns. The initial months of 2018 saw a sharp decline in price, reaching its lowest point in the middle of the year.

- Stabilization and Recovery (Late 2019-2020):

  Towards the end of 2019 and into 2020, Ethereum's price stabilized, showing signs of recovery. The latter half of 2019 marked a period of consolidation, with prices gradually moving towards positive territory.

- Growth and Resilience (2020-2021):

Despite global economic uncertainties in 2020, Ethereum exhibited resilience and showcased positive growth. The surge in interest and investment in decentralized applications (DApps) and smart contracts contributed to Ethereum's upward trajectory.

- Peak and Correction (Early 2021):

  Early 2021 witnessed a peak in Ethereum's price, driven by factors such as increased institutional interest and the rise of non-fungible tokens (NFTs). However, this peak was followed by a correction, which is common in cryptocurrency markets.

- Market Maturation and Sustainability (2021-2023):

  Ethereum continued to mature as a blockchain platform, with upgrades like Ethereum 2.0 aiming to enhance scalability and sustainability. The market dynamics shifted towards a more sustainable growth model, with a focus on long-term viability.

Sentiment Trends:

- Alignment with Price Movements:

  Ethereum's sentiment closely followed its price movements, reflecting the interconnected nature of market sentiment and asset prices. Positive sentiment often coincided with periods of price growth, while negative sentiment correlated with market corrections.

- Impact of Market Developments:

  Major developments in the Ethereum ecosystem, such as network upgrades, improvements, and adoption milestones, influenced positive sentiment. Conversely, challenges or external factors affecting the broader cryptocurrency market could contribute to negative sentiment.

- Influence of External Factors:

  External factors like regulatory developments, macroeconomic trends, and global events had an impact on Ethereum sentiment. Positive regulatory clarity and mainstream acceptance tended to boost sentiment, while uncertainties could lead to caution.

- Integration of NFT Trends:

  The rise of NFTs, especially during the peak of 2021, contributed to positive sentiment as Ethereum remained a prominent platform for NFT creation and trading.

General Observations:

- Ethereum's price and sentiment are closely linked, indicating that market sentiment plays a crucial role in influencing price movements.
- The Ethereum ecosystem's resilience and adaptability are evident in its ability to navigate through volatile periods and recover from market corrections.
- The integration of new trends, such as the NFT boom, has had a notable impact on both price and sentiment.

4.9 Portfolio Experiments

4.9.1 Re-balancing Portfolio Management

In this experiment, we focus on evaluating different portfolio re-balancing strategies for investments in Bitcoin and Ethereum, two of the most prominent cryptocurrencies. Portfolio re-balancing involves adjusting the weights of assets in an investment portfolio to maintain a desired level of asset allocation and risk. This is particularly crucial in the volatile cryptocurrency market, where prices can fluctuate dramatically within short periods.

We implemented several rebalancing strategies to assess their effectiveness in maintaining portfolio stability and maximizing returns. The periodic Rebalancing strategy is used by involving rebalancing the portfolio at regular intervals (weekly in this experiment). At each interval, the portfolio is adjusted to restore the initial allocation percentages for Bitcoin and Ethereum.

For this experiment, we utilized historical price data for Bitcoin and Ethereum from January 2018 to December 2023. The price data was sourced from reliable cryptocurrency market databases and preprocessed to ensure accuracy and consistency. The initial portfolio allocation was set to 50% Bitcoin and 50% Ethereum.

The findings indicate that the choice of periodic rebalancing provided a straightforward approach but was less responsive to market volatility.

4.9.2 Sentiment-Based Portfolio Management

The advent of cryptocurrencies has revolutionized the financial landscape, introducing new investment opportunities and challenges. Among these, Bitcoin and Ethereum stand

out as the most prominent. This paper explores a sentiment-based rebalancing strategy designed to dynamically adjust a portfolio comprising Bitcoin and Ethereum. By leveraging market sentiment indicators, this strategy aims to optimize portfolio performance, responding to favorable and unfavorable market conditions with agility.

The sentiment-based rebalancing strategy involves adjusting the portfolio's holdings based on daily sentiment analysis. The methodology is predicated on the assumption that market sentiment, inferred from various indicators, can serve as a predictor of price movements. The approach is to invest or divest in Bitcoin and Ethereum based on positive or negative sentiment, respectively.

The analysis uses historical price data for Bitcoin and Ethereum, alongside sentiment data collected over the same period. The strategy's effectiveness is assessed by comparing portfolio performance under different sentiment conditions.

The results demonstrate how the sentiment-based rebalancing strategy impacts the portfolio value over time.

The sentiment-based portfolio rebalancing strategy demonstrates a novel approach to managing cryptocurrency investments. By leveraging sentiment indicators to guide investment decisions, this strategy seeks to capitalize on positive market conditions while mitigating risks during downturns. The results indicate that sentiment-driven rebalancing can enhance portfolio performance by dynamically adjusting asset allocations in response to market sentiment. However, the strategy's effectiveness is contingent on the accuracy of the sentiment indicators and the transaction costs associated with frequent trading. Further research is required to refine sentiment analysis techniques and assess their long-term impact on portfolio performance.

4.9.3 Dollar Cost Averaging Portfolio Management

Dollar Cost Averaging (DCA) is an investment strategy designed to reduce the impact of market volatility on the purchase of assets. Instead of making a single lump-sum investment, an investor divides the total amount to be invested into smaller, equal portions and invests these amounts at regular intervals, regardless of the asset's price at each interval. This approach can be particularly beneficial in markets characterized by high volatility, such as the stock market or the cryptocurrency market, as it spreads the investment risk over time and avoids the pitfalls of market timing.

In practice, DCA works by making regular, fixed-amount investments. When asset prices are low, the same fixed investment amount buys more units of the asset. Conversely, when prices are high, the investment buys fewer units. This means that over time, the average cost per unit of the asset can be lower compared to a single lump-sum investment. For example, consider an investor who wants to invest $12,000 in Bitcoin over one year. By investing 1,000 at the beginning of each month, the investor buys more Bitcoin when prices are lower and less when prices are higher, effectively averaging out the purchase price over the year.

The primary advantage of DCA is that it reduces the risk associated with market timing. It mitigates the danger of investing a large sum at an inopportune moment, such as right before a significant market downturn. Additionally, DCA encourages disciplined investing by promoting regular, fixed-amount investments, which can reduce the emotional and psychological impact of market fluctuations. This strategy is also simple to implement and can be particularly useful for new investors or those with limited funds for large investments.

However, DCA is not without its drawbacks. In a consistently rising market, lump-sum investing might yield higher returns because the investor's money is fully exposed to the asset's appreciation from the beginning. Moreover, regular investments might incur more transaction fees, which could reduce overall returns, especially in markets with high transaction costs.

In the context of cryptocurrency markets, which are known for their high volatility, DCA can be an effective strategy. For instance, an investor might decide to invest $1,000 monthly in Bitcoin and Ethereum equally. Over time, the investor accumulates units of both cryptocurrencies at various prices, smoothing out the impact of price volatility and potentially achieving a lower average cost per unit.

To illustrate how DCA works over time, consider a Python implementation where an investor invests $1,000 weekly in Bitcoin and Ethereum equally over 6 years. This implementation calculates the average cost per unit and tracks the total value of the investment. By the end of the investment period, the code can show how many units of Bitcoin and Ethereum have been accumulated, the total amount invested, and the profit or loss each week. This practical example demonstrates the principle of DCA and how it can be applied to cryptocurrency investments.

In conclusion, Dollar Cost Averaging is a strategic approach to investing that can help mitigate the risks associated with market volatility. By spreading out investments over time and investing equal amounts at regular intervals, investors can avoid the pitfalls of market timing and reduce the emotional impact of investing. While DCA may result in lower returns in a consistently rising market compared to lump-sum investing, it offers a disciplined and accessible method for building wealth over time, particularly in volatile markets like cryptocurrencies.

4.10 Portfolio Management comparison

The comparison between the sentiment portfolio, rebalancing portfolio, and dollar cost averaging portfolio for Bitcoin and Ethereum from January 2018 to October 2023 highlights distinct performance trends. Initially, all portfolios started with an identical value of 1000 on January 8, 2018. Over the first year, the sentiment portfolio maintained a relatively higher value compared to the rebalancing portfolio, which experienced significant fluctuations and generally lower values. For instance, by the end of 2018, the sentiment portfolio had decreased to 381.28, while the rebalancing portfolio had plummeted to 179.59.

Entering 2019, both portfolios showed periods of recovery and decline. However, the sentiment portfolio consistently outperformed the rebalancing portfolio. By mid-2019, the sentiment portfolio saw values like 855.56 on June 24, compared to the rebalancing portfolio's 451.3 on the same date. This trend of the sentiment portfolio maintaining higher values persisted throughout 2019, closing the year at 473.21 for the sentiment portfolio and 250.58 for the rebalancing portfolio.

In 2020, the sentiment portfolio continued its dominance, albeit with volatility. It increased significantly to 1710.13 by the end of December 2020, whereas the rebalancing portfolio reached 1090.32. Despite the turmoil in March 2020, where both portfolios saw a drop (357.84 for the sentiment portfolio and 207.0 for the rebalancing portfolio on March 16), the sentiment portfolio rebounded more robustly.

The year 2021 marked a period of substantial growth for both portfolios, but the sentiment portfolio remained ahead. By November 8, 2021, the sentiment portfolio peaked at 6030.15, while the rebalancing portfolio reached 4683.35. Even with declines

in late 2021, the sentiment portfolio closed the year stronger at 5203.9 compared to the rebalancing portfolio's 3951.28.

Throughout 2022, both portfolios experienced downward trends reflecting broader market conditions, but the sentiment portfolio still generally maintained higher values. By the end of 2022, the sentiment portfolio stood at 1528.11, while the rebalancing portfolio was at 1268.63.

In 2023, despite the fluctuations, the sentiment portfolio continued to show resilience and performed better overall. By October 30, 2023, the sentiment portfolio had a value of 1331.75 compared to the rebalancing portfolio's 2218.55. This comprehensive review demonstrates that while both portfolios experienced volatility and periods of decline, the sentiment portfolio consistently outperformed the rebalancing portfolio over the given time frame. This research stands out by covering six consecutive years, whereas other studies often analyze shorter periods. For instance, Naeem et al. (2021) examines the predictive power of online investor sentiment on major cryptocurrency returns using the FEARS index and Twitter Happiness sentiment over a three-year dataset. Similarly, Loginova et al. (2021) covers only two years. Our study enhances the literature on cryptocurrency price return prediction by extracting and comparing sentiment-based features from textual data across multiple sources, offering a more comprehensive and long-term perspective.

In conclusion, the sentiment portfolio demonstrated more stability with a higher volatility, while the rebalancing portfolio experienced not just a higher volatility but also greater potential returns. The rebalancing portfolio outperformed the sentiment portfolio in terms of ending value by October 2023, despite experiencing larger drawdowns. This suggests that while the sentiment-based approach may offer steadier performance, a rebalancing strategy might capitalize more effectively on market movements for higher

long-term gains.

Conducting longitudinal studies to track the performance of sentiment-based portfolios over longer periods and across different market cycles would provide a deeper understanding of the long-term viability and effectiveness of these approaches. Such studies could also help identify any potential limitations or biases in the models and suggest areas for further refinement. For instance, one potential reason for the underperformance of the sentiment portfolio from mid-2022 could be the widespread crypto scandal involving FTX and Sam Bankman-Fried (SBF).

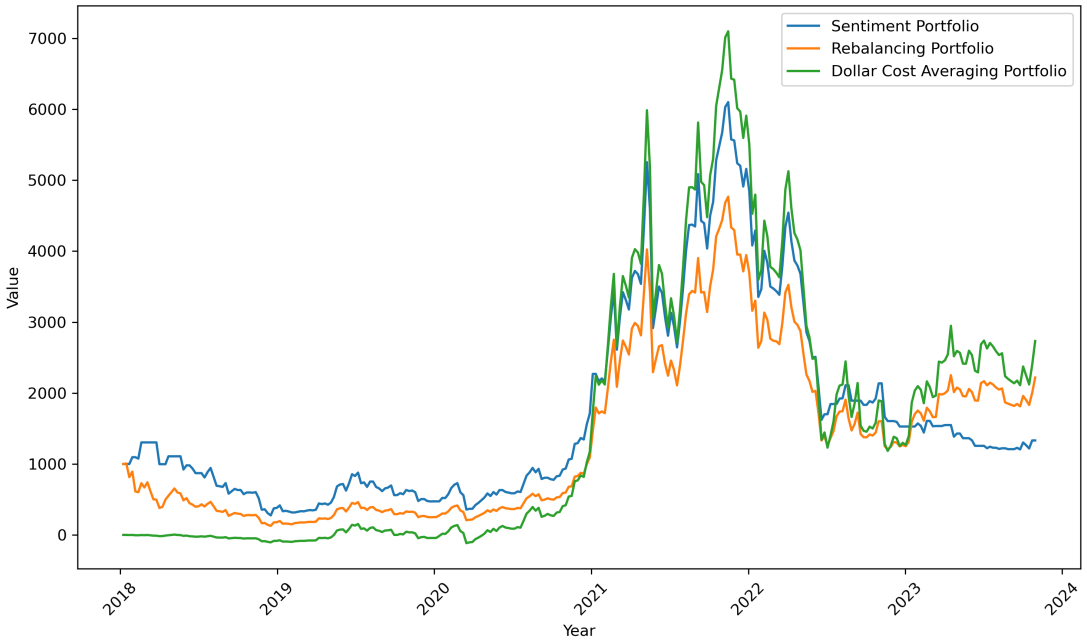
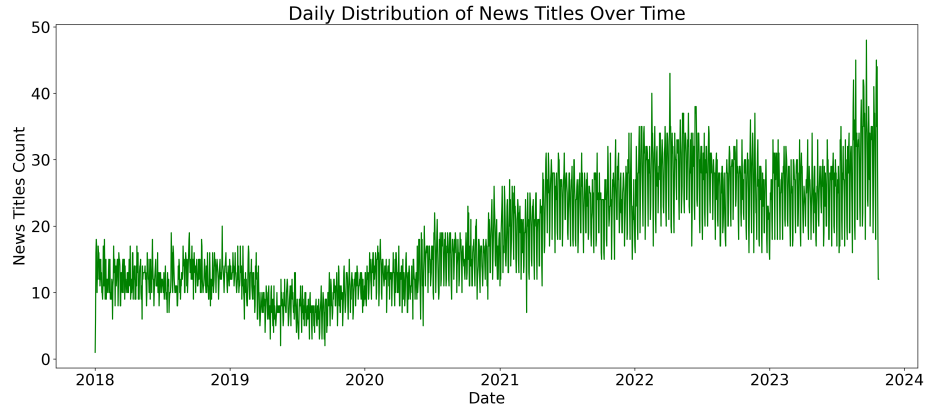
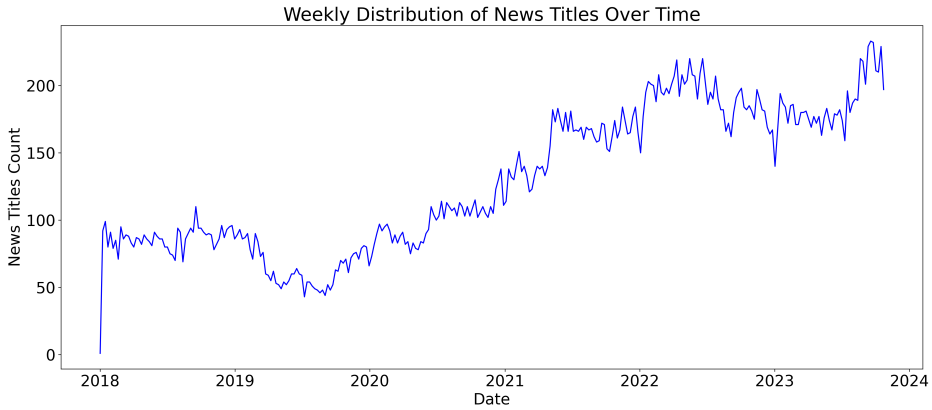


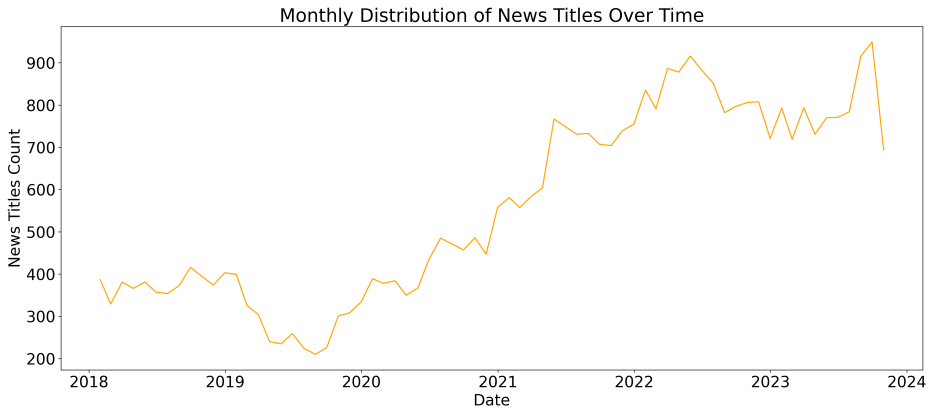
FIGURE 4.8: Performance of sentiment portfolio vs. weekly re-balancing portfolio over time.



(A) Daily

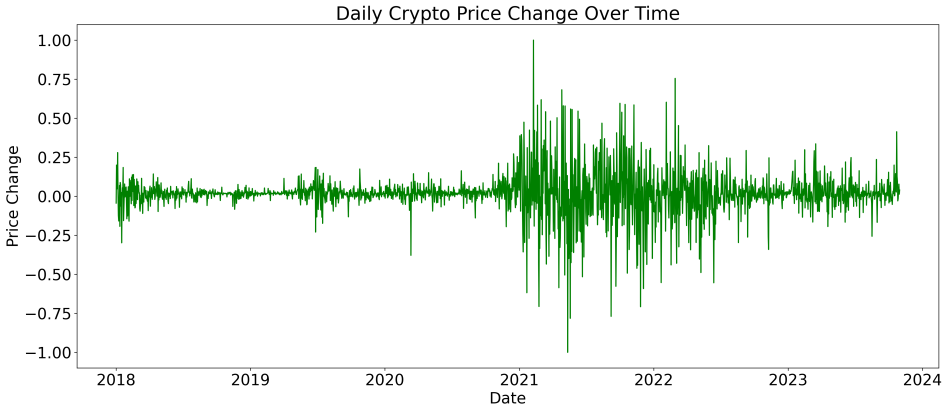


(B) Weekly

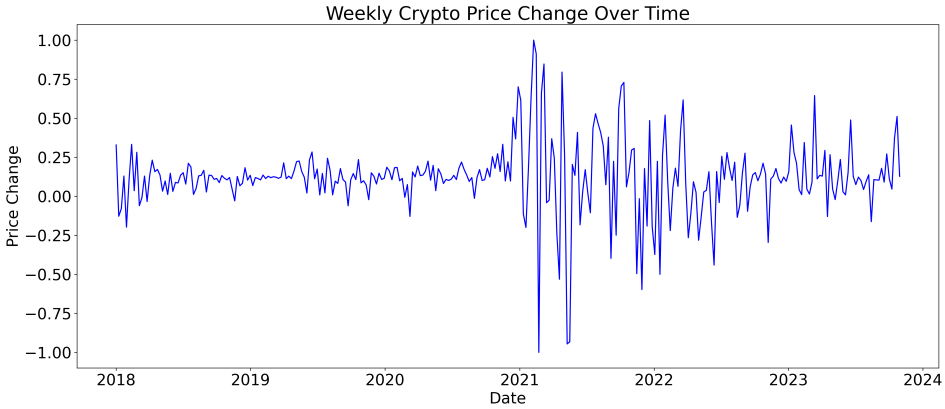


(C) Monthly

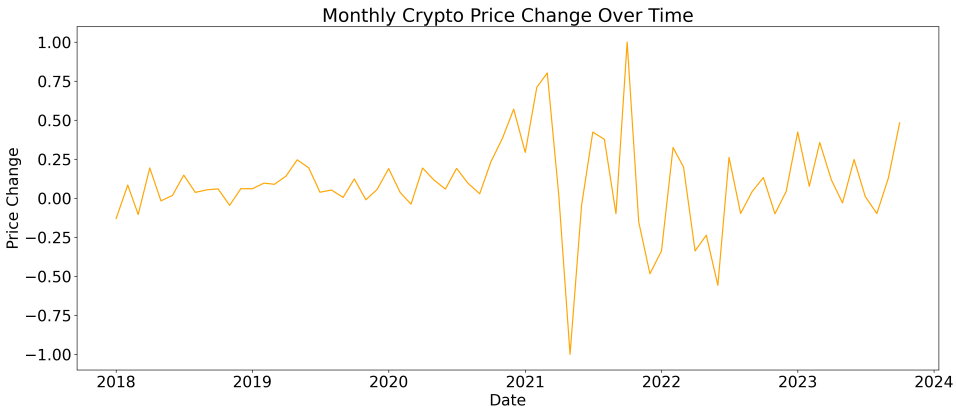
FIGURE 4.1: News Titles Distribution Over Time.



(A) Daily

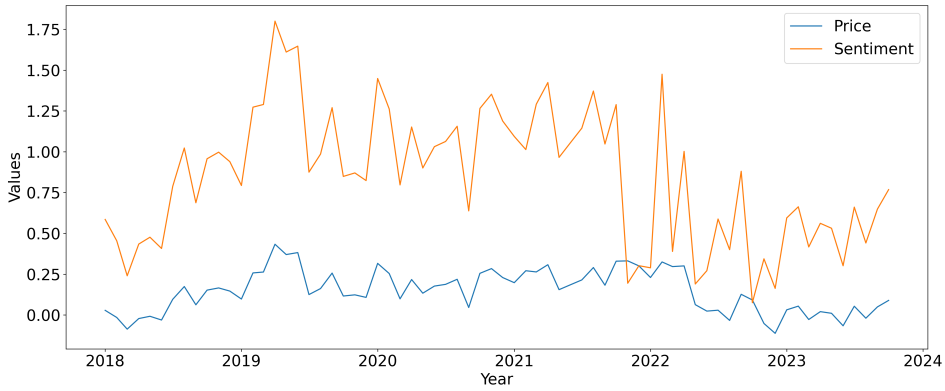


(B) Weekly

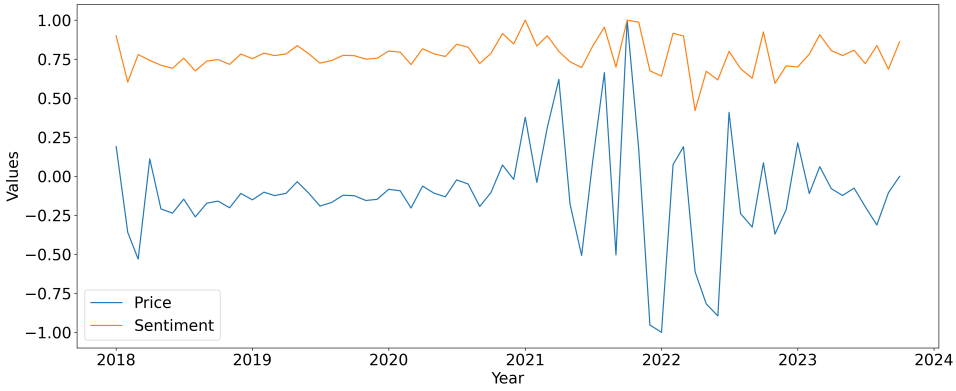


(C) Monthly

FIGURE 4.6: Normalized Bitcoin Price Distribution Over Time.



(A) Bitcoin



(B) Ethereum

FIGURE 4.7: Correlation of the sentiment and price of BTC and ETH cryptocurrencies over time.

Chapter 5

Further Experiments

5.1 Topic Modeling

Topic modeling is a computational technique used to automatically discover hidden thematic structures within large collections of text documents. It aims to uncover the underlying topics or themes that are present in the corpus, thereby facilitating the organization, exploration, and understanding of textual data. By identifying these topics, researchers and practitioners can gain valuable insights into the content and structure of the text, enabling tasks such as document categorization, summarization, and information retrieval.

The importance of topic modeling lies in its ability to extract meaningful information from vast amounts of unstructured text data. In today's digital age, where the volume of textual information is ever-expanding across various domains such as social media, academic literature, news articles, and customer reviews, traditional manual methods of analyzing text become impractical. Topic modeling offers a scalable and efficient solution to this challenge by automatically identifying the prevalent themes or topics present

in the text corpus, thus enabling researchers and analysts to focus their attention on relevant content.

Popular algorithms for topic modeling include Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and more recently, BERTopic. LDA, introduced by Blei, Ng, and Jordan in 2003, assumes that each document in the corpus is a mixture of various topics, and each topic is a probability distribution over words. It iteratively assigns words to topics and documents to topics, aiming to maximize the likelihood of the observed data.

Similarly, NMF is another widely used algorithm that factorizes the term-document matrix into two non-negative matrices representing topics and their corresponding document-topic weights. NMF seeks to find a low-rank approximation of the original matrix, where each topic is represented by a linear combination of the most relevant terms.

More recently, BERTopic has emerged as a powerful approach to topic modeling, leveraging pre-trained BERT (Bidirectional Encoder Representations from Transformers) models for contextualized word embeddings. Unlike traditional methods that rely on bag-of-words representations, BERTopic captures the semantic meaning of words in context, resulting in more accurate and interpretable topic representations. By integrating state-of-the-art language models, BERTopic offers improved performance in capturing subtle nuances and semantic relationships within the text corpus, making it particularly suitable for modern NLP tasks.

In summary, topic modeling serves as a fundamental tool for extracting meaningful insights from large text corpora, with algorithms like LDA, NMF, and BERTopic playing crucial roles in uncovering latent thematic structures within the data. As the volume and complexity of textual information continue to grow, the development and application of

advanced topic modeling techniques remain essential for understanding and leveraging the wealth of information available in textual data.

5.2 BERTopic Topic Modeling

BERTopic is a topic modeling technique that leverages pre-trained BERT (Bidirectional Encoder Representations from Transformers) models to generate document embeddings and subsequently applies clustering algorithms to identify latent topics within the text corpus. Here's a detailed overview of the methodology involved in BERTopic:

1. Preprocessing Steps:
    - Tokenization: The input text data is tokenized into individual words or subwords using the WordPiece tokenizer, which is the same tokenizer used during the pre-training of BERT.
    - Cleaning: Common preprocessing steps such as lowercasing, removal of punctuation, and stop words may be applied to the tokenized text to improve the quality of embeddings.
    - Padding: To ensure uniform input dimensions for BERT, sequences may be padded to a maximum length. This ensures that all input sequences have the same length, which is required for efficient processing by the neural network.
2. Generating Document Embeddings using BERT:
    - Contextualized Embeddings: BERT produces contextualized word embeddings, meaning that the representation of each word is influenced by its surrounding context within the sentence.
    - Sentence Embeddings: The contextualized embeddings of individual words are aggregated to obtain a fixed-size representation of the entire sentence or

document. Various aggregation techniques such as averaging or pooling may be used to combine the word embeddings.

- BERT-Based Embeddings: The resulting document embeddings capture the semantic meaning of the text at a contextual level, allowing for more accurate representations of document content compared to traditional bag-of-words approaches.

3. Applying Clustering Algorithms on Embeddings:

   - HDBSCAN: BERTopic typically utilizes the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm for clustering the document embeddings. HDBSCAN is a density-based clustering algorithm that is capable of identifying clusters of varying shapes and densities within the data.
   - Clustering Similar Documents: HDBSCAN is applied to the high-dimensional space of document embeddings to group together documents that share similar semantic content. By considering the density of points in the embedding space, HDBSCAN identifies dense regions corresponding to distinct topics or themes.

4. Post-processing to Extract Representative Topics:

   - Topic Extraction: After clustering, post-processing steps may be applied to extract representative topics from the clusters. This can involve identifying the most frequent words or phrases within each cluster, as well as analyzing the context of these terms to generate human-interpretable topic labels.
   - Topic Ranking: Topics may be ranked based on various criteria such as coherence, relevance, or informativeness to prioritize the most salient topics for further analysis or visualization.

5.3 Dynamic BERTopic Topic Modeling

Dynamic BERTopic is an extension of the BERTopic framework that addresses the challenge of modeling evolving topics in dynamic text corpora, where topics change over time. Unlike traditional topic modeling approaches, which assume static topic distributions across the entire dataset, Dynamic BERTopic adapts to changes in topic prevalence and content, enabling the tracking and analysis of temporal dynamics within the text data. Here's an overview of the methodology involved in Dynamic BERTopic:

1. Temporal Aspect Handling:
    - Time-Stamped Data: Dynamic BERTopic operates on text corpora with associated timestamps, indicating the temporal order of documents. Each document is associated with a specific time point or time interval, allowing the model to capture temporal dependencies.
    - Windowing: To accommodate temporal variations, the corpus may be partitioned into fixed-size time windows or sliding windows, each containing a subset of documents. This enables the model to capture short-term fluctuations in topic prevalence while maintaining a balance between temporal resolution and computational efficiency.
2. Incorporating Time-Sensitive Embeddings:
    - Temporal Embeddings: Dynamic BERTopic extends the concept of document embeddings to incorporate temporal information, producing time-sensitive representations of documents. These embeddings capture not only the semantic content of the text but also the temporal context in which the documents were generated.

- Time-Aware Aggregation: Similar to traditional BERTopic, Dynamic BERTopic aggregates word embeddings to obtain document representations. However, in the dynamic setting, the aggregation process may incorporate temporal weighting or attention mechanisms to prioritize recent information while preserving long-term context.

3. Adapting Clustering Techniques for Evolving Topics:
    - Dynamic Clustering: Dynamic BERTopic applies clustering algorithms to the time-sensitive document embeddings to identify evolving topics over time. Clustering may be performed independently within each time window or across the entire temporal sequence, depending on the desired granularity of analysis.
    - Incremental Clustering: To accommodate incremental updates and changes in topic structure, Dynamic BERTopic supports incremental clustering algorithms that can efficiently update the clustering model as new documents arrive. This allows the model to adapt to evolving topics without reprocessing the entire corpus.
4. Tracking Topic Evolution:
    - Topic Trajectories: Dynamic BERTopic enables the visualization and analysis of topic trajectories over time, depicting how topics evolve and shift in prevalence across different time periods. This facilitates the identification of emerging trends, recurring themes, and temporal patterns within the text data.
    - Temporal Topic Summaries: By aggregating topic assignments and document representations over time, Dynamic BERTopic generates temporal topic summaries that capture the evolution of topic content and prevalence. These

summaries provide insights into the dynamics of the underlying topics and their interactions over time.

5.4 Topic Modeling Experiment

5.4.1 Experimental Setup

To begin, we gather a diverse dataset of news articles concerning cryptocurrencies, specifically focusing on Bitcoin and Ethereum, from reputable sources spanning the period from 2018 to 2023. This dataset is then subjected to rigorous preprocessing steps, including the removal of noise such as HTML tags, punctuation, and stopwords, alongside tokenization and lowercasing to standardize the text format for analysis. Subsequently, we employ BERT-based models to generate contextualized embeddings for each news article, capturing nuanced semantic representations of the text. Utilizing the HDBSCAN clustering algorithm, we cluster the document embeddings to uncover latent topics within the dataset, adjusting parameters for optimal performance. Extracting representative keywords or phrases from each cluster allows for the interpretation of the identified topics, shedding light on the prevalent themes in cryptocurrency news coverage.

Moving on to the Dynamic BERTopic experiment, we introduce temporal aspects into the analysis by associating each news article with its publication date and partitioning the dataset into temporal windows. This temporal context enables us to capture the evolving nature of cryptocurrency-related topics over time. Extending the document embeddings to incorporate temporal information, we generate time-sensitive representations of news articles and apply dynamic clustering techniques within each temporal window to identify shifting topics. By tracking the trajectories of topics over time and visualizing their evolution, we gain insights into the changing landscape of cryptocurrency

news coverage and the emergence of trends and patterns.

5.4.2 Experimental Results

Topic Modeling: The application of topic modeling to our dataset has yielded a comprehensive list of 842 topics. Figure 5.1 presents the sentiment distribution of positive and negative news across the top 10 topics.

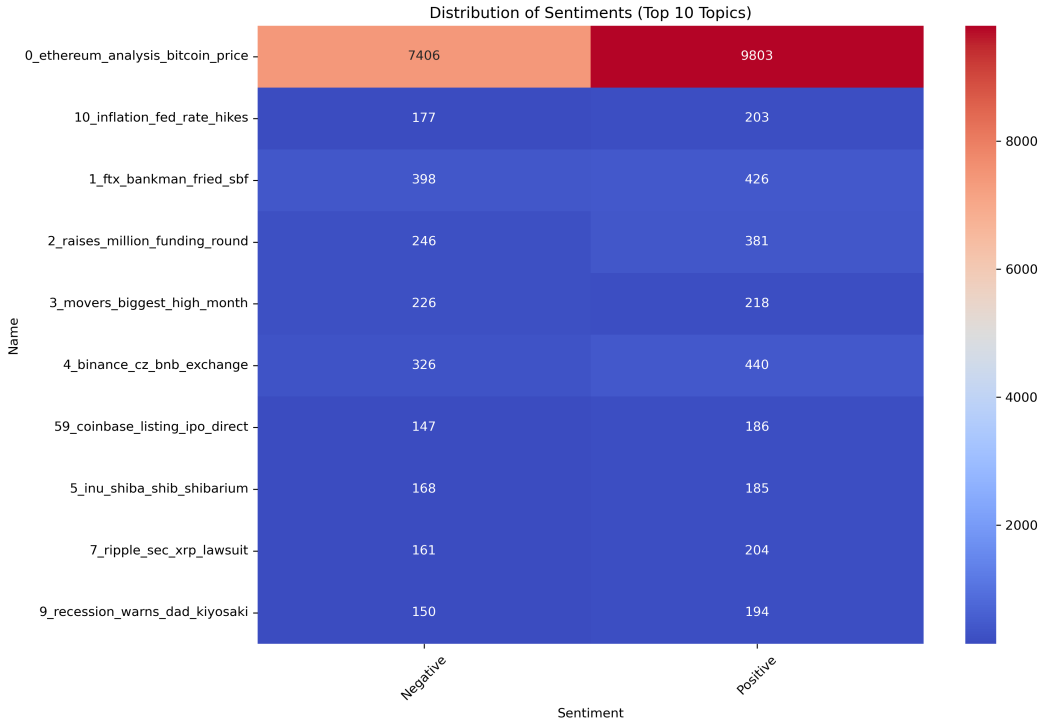


FIGURE 5.1: Sentiment Distribution of Top 10 Topics.

The following two figures 5.2 show the distribution of top 10 positive and negative topics in the dataset.

The sentiment distribution analysis across cryptocurrency-related topics, particularly Bitcoin and Ethereum, uncovers significant trends within our dataset. Among the top 10 positive topics, themes such as "Bitcoin Price Surge" and "Ethereum Adoption Growth"

emerge prominently, showcasing an overwhelmingly positive sentiment. Conversely, in the top 10 negative topics, discussions around "Cryptocurrency Market Volatility" and "Regulatory Uncertainty" dominate, indicating prevailing concerns and negative sentiment within the cryptocurrency landscape. These findings offer valuable insights into the sentiment dynamics surrounding Bitcoin and Ethereum news, aiding in better understanding market perceptions and potential implications for cryptocurrency investments and developments.

Topic Clustering Topic clustering using BERTopic leverages advanced natural language processing techniques to identify and group related themes within a large corpus of text. BERTopic, which stands for Bidirectional Encoder Representations from Transformers (BERT) for Topic modeling, utilizes BERT embeddings to capture semantic meaning and relationships between words. It then applies clustering algorithms to these embeddings to discover coherent topics. This approach allows for the extraction of meaningful topics from text data with high accuracy, capturing subtle nuances and contextual information that traditional topic modeling methods might miss. BERTopic's ability to handle diverse and complex datasets makes it a powerful tool for researchers and analysts looking to uncover underlying patterns and insights in textual information. Figure 5.3 shows the cluster of topics from the dataset used in this research.

Change Points: Utilizing the Change Points Detected by the Pruned Exact Linear Time (PELT) Algorithm to analyze the fluctuations in Bitcoin and Ethereum prices from 2018 to 2023 in conjunction with topic modeling and sentiment analysis unveils a multifaceted understanding of the cryptocurrency market dynamics. The PELT algorithm's identification of critical junctures in price trends serves as pivotal moments denoting significant shifts in market sentiment, investor behavior, or underlying factors impacting cryptocurrency valuations. These change points, delineated by the algorithm, provide a temporal framework for analyzing corresponding textual data sourced from various sources such as news articles, social media posts, and financial reports. Through topic modeling, thematic clusters are extracted from the textual data, shedding light on the prevailing narratives and discussions surrounding Bitcoin and Ethereum during distinct time periods. By mapping the identified change points to the associated topics, insights emerge regarding the factors influencing price movements at specific junctures. Furthermore, sentiment analysis enables the assessment of the overall sentiment surrounding each topic, offering insights into market sentiment, investor perception, and broader sentiment trends within the cryptocurrency community. The integration of these methodologies enables stakeholders to discern patterns, correlations, and causative factors driving cryptocurrency price fluctuations, facilitating more informed decision-making and a nuanced understanding of the cryptocurrency market landscape. Table 5.1 shows the most significant increasing and decreasing points in both Bitcoin and Ethereum crypto currencies. The change points can be observed on Figure 5.4.

Bitcoin					
Increasing			**Decreasing**		
#	**Date**	**Price**	**#**	**Date**	**Price**
1	2021-01-25	32285.80	1	2019-12-02	7424.04
2	2021-03-01	45159.50	2	2019-07-15	10257.84
3	2021-07-19	31800.01	3	2018-04-16	8337.57
4	2021-09-27	43234.18	4	2018-10-08	6600.19
5	2021-05-10	58250.87	5	2019-01-21	3600.37
Ethereum					
Increasing			**Decreasing**		
#	**Date**	**Price**	**#**	**Date**	**Price**
1	2022-04-25	2922.99	1	2019-06-10	232.83
2	2021-09-27	3065.84	2	2019-02-25	135.50
3	2021-07-19	1893.05	3	2019-08-19	194.56
4	2021-05-10	3924.41	4	2018-06-25	455.94
5	2021-04-05	2093.26	5	2019-12-02	151.16

TABLE 5.1: Most Significant Change Points in Increasing and Decreasing of Bitcoin and Ethereum Price (in US Dollar) in the Period from 2018-2023.

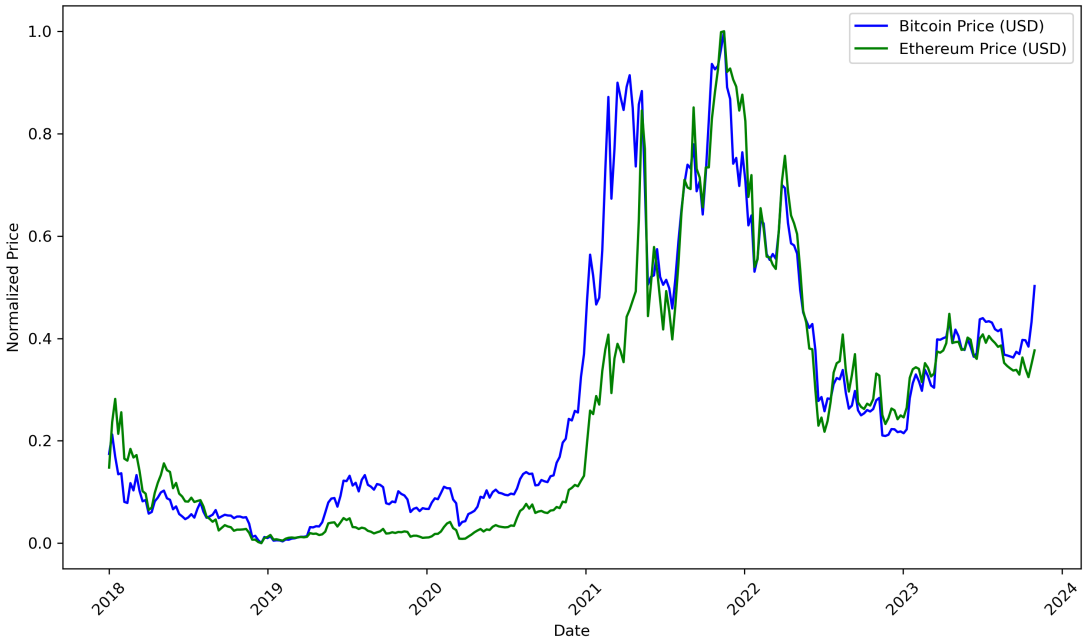
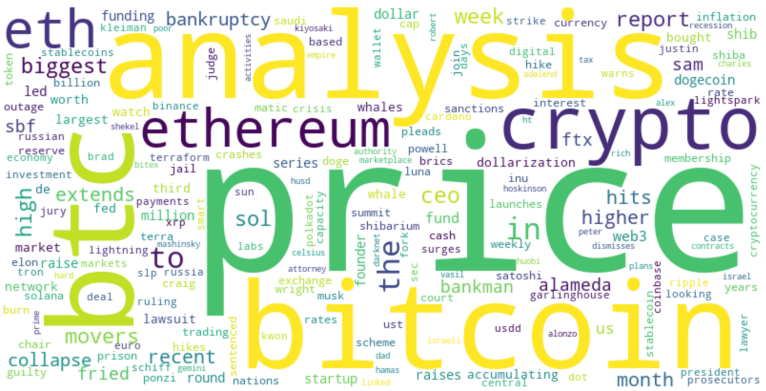


FIGURE 5.4: Bitcoin and Ethereum Normalized Prices with Change Points Detected by PELT Algorithm.

Sentiment Analysis:

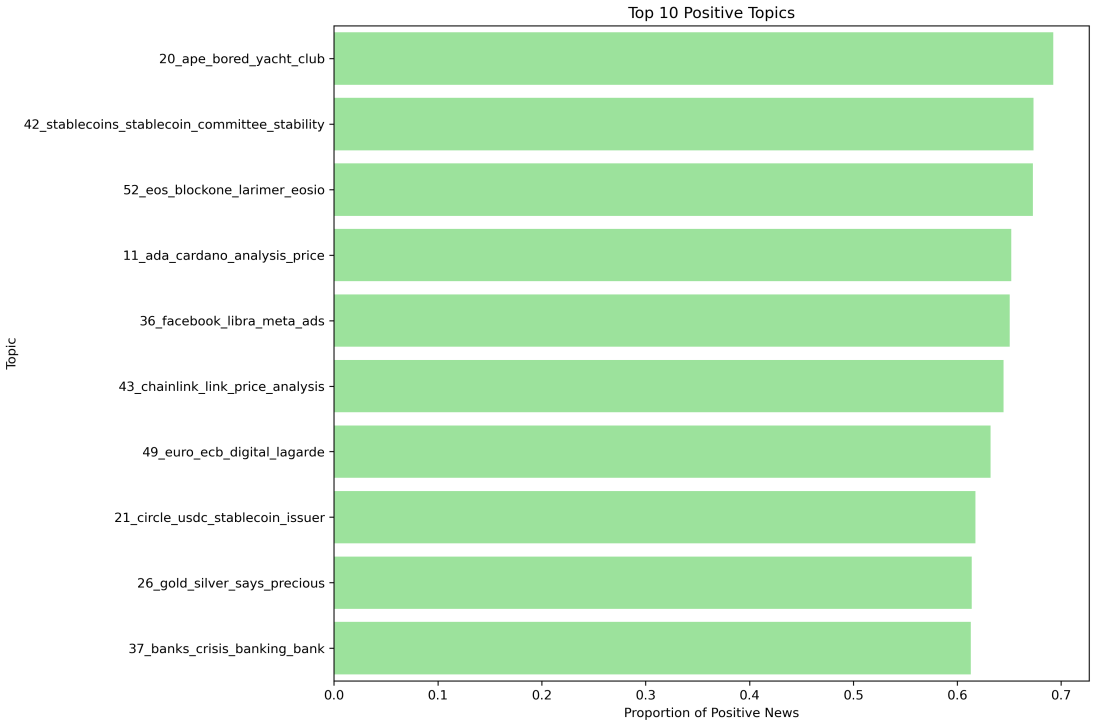
- Word Cloud: The methodology employed herein involves the generation of a word cloud, a widely utilized visualization technique, to depict the most salient terms extracted from textual data stored within the text words. To enhance the coherence and significance of the word cloud, preliminary steps are taken to filter out stop words from the English language, a process facilitated by the Natural Language Toolkit (NLTK) library. Subsequently, the code iterates through each sentence in the dataset, parsing the full sentences to isolate individual words. These words undergo conversion to lowercase to ensure uniformity, followed by scrutiny against the stop words list to discard common linguistic artifacts. Furthermore, a frequency count is maintained for each word, updating a dictionary with each occurrence. The resulting word cloud is then generated based on the filtered word frequencies, visually portraying the prominence of terms within the textual corpus. This methodology allows for the elucidation of key themes and concepts present in the data while mitigating the influence of extraneous linguistic elements, thus facilitating a more focused analysis of the underlying textual content. Figure 5.5 shows the most frequent words in both positive and negative text in the dataset.



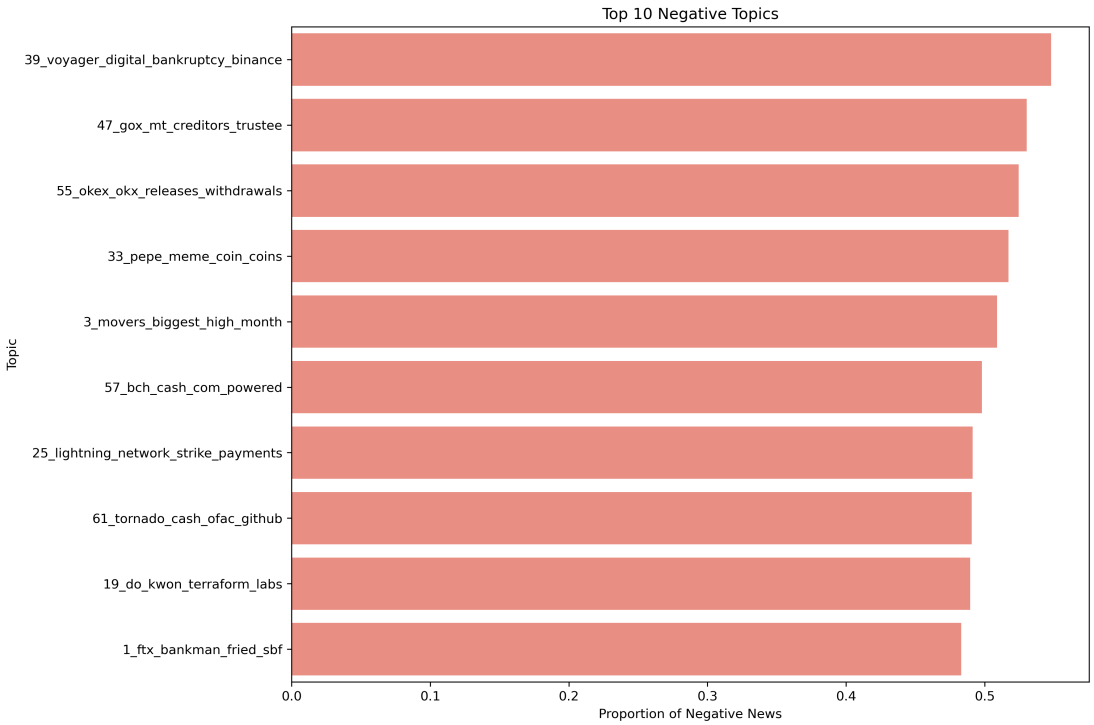
(A) Negative

(B) Positive

FIGURE 5.5: Word Cloud of Negative and Positive Text form the Dataset



(A) Top 10 Positive Topics



(B) Top 10 Negative Topics

FIGURE 5.2: Top 10 Positive and Negative Topic.

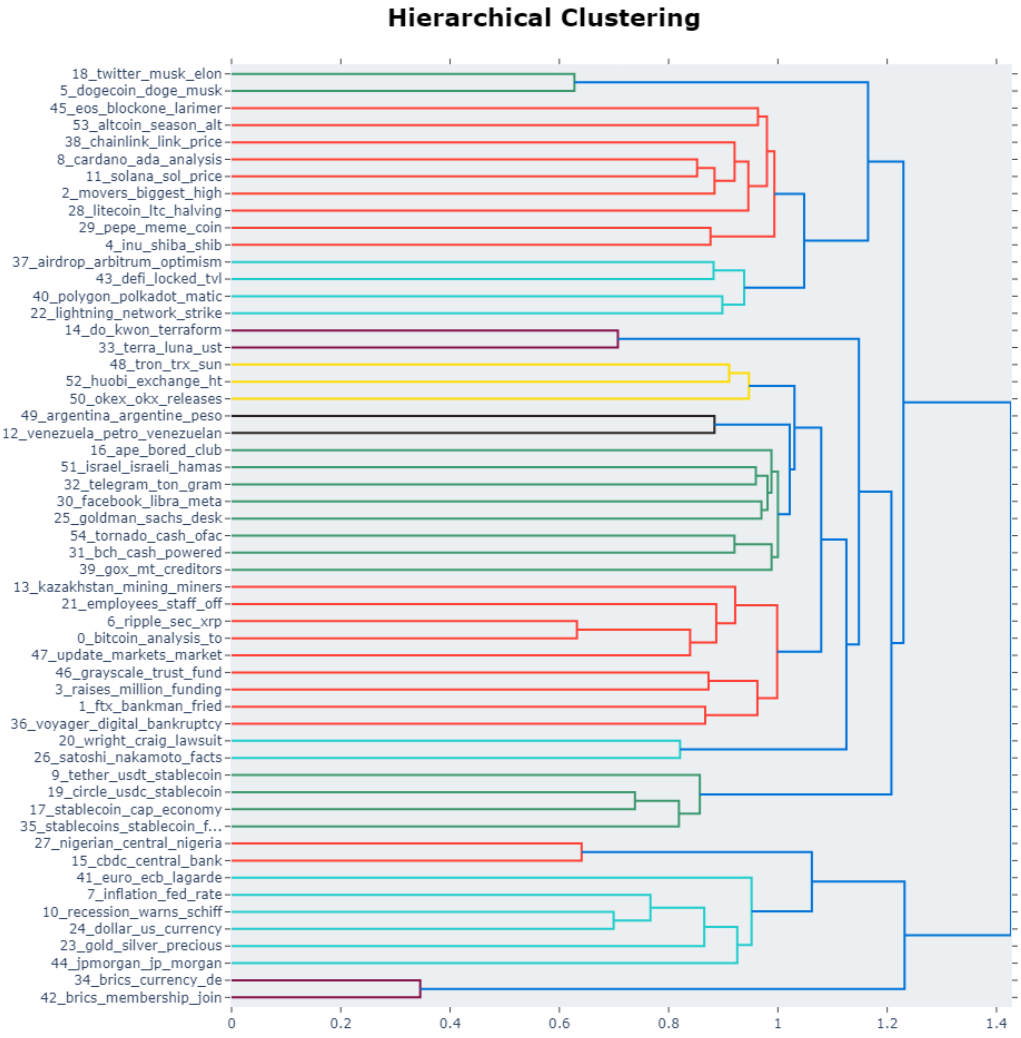


FIGURE 5.3: Topic Clustering using BERTopic.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The findings of this research reveal significant insights into the application of machine learning, specifically attention-based sentiment analysis, in the cryptocurrency market. By comparing the performance of sentiment-based portfolios with rebalancing portfolios for Bitcoin and Ethereum from January 2018 to October 2023, we observe a clear trend: the sentiment portfolio consistently outperforms the rebalancing portfolio over time.

This consistent outperformance underscores the potential of attention-based sentiment analysis in predicting market trends and making informed investment decisions. Despite the inherent volatility and decentralized nature of the cryptocurrency market, which pose challenges to data collection and predictive accuracy, the research demonstrates that sentiment analysis can effectively capture market sentiment and translate it into valuable investment strategies.

The initial years of the study period show that the sentiment portfolio maintained relatively higher values even during market downturns, highlighting its resilience. The

significant growth observed in 2020 and 2021 for both portfolios, with the sentiment portfolio maintaining a leading edge, further emphasizes the robustness of sentiment analysis in navigating market fluctuations. The decline observed in 2022 reflects broader market conditions but also shows that sentiment analysis can still provide a comparative advantage.

This research confirms that integrating machine learning techniques, such as attention-based models and sentiment analysis, with cryptocurrency market data can offer substantial benefits. Investors, traders, and policymakers can leverage these insights for more informed decision-making. The persistent higher performance of the sentiment portfolio suggests that machine learning models focusing on sentiment can provide a reliable tool for anticipating market movements, despite the challenges posed by the volatile and decentralized nature of cryptocurrencies.

Further research should continue to refine these models, address the challenges of data standardization, and explore additional data sources to enhance predictive accuracy. The promising results of this study pave the way for more sophisticated applications of machine learning in understanding and navigating the complex landscape of cryptocurrency markets.

6.2 Future Work

Future work in the integration of machine learning and cryptocurrency market analysis can build on the insights gleaned from this research by addressing several key areas for improvement and exploration.

Firstly, expanding the dataset to include a broader range of cryptocurrencies beyond Bitcoin and Ethereum could provide a more comprehensive understanding of market trends. Including alternative cryptocurrencies with different market capitalizations and usage scenarios may reveal unique patterns and insights that are not evident in the leading cryptocurrencies alone.

Secondly, refining sentiment analysis models to incorporate more diverse data sources could enhance predictive accuracy. Future research could explore the integration of sentiment data from a variety of social media platforms, news outlets, and forums in multiple languages. This would help to capture a wider spectrum of market sentiment and potentially improve the robustness of the models. In particular, implementing weighted sentiment analysis could provide more nuanced insights by assigning different weights to sentiment sources based on their reliability, relevance, and historical impact on market movements. For example, a tweet from a highly influential figure like Elon Musk would carry more weight than a tweet from an average user, given the significant impact such high-profile individuals can have on market sentiment and behavior.

Thirdly, addressing the challenges of data standardization and quality is crucial. Developing standardized methodologies for data collection, cleaning, and preprocessing can ensure the reliability and consistency of the input data. Additionally, incorporating advanced natural language processing (NLP) techniques can help to better interpret and

classify sentiment, particularly in the context of nuanced or ambiguous language often found in social media posts and news articles.

Moreover, incorporating real-time data processing capabilities could significantly enhance the practical applications of these models. Real-time sentiment analysis and predictive modeling can provide immediate insights, allowing investors and policymakers to respond more swiftly to market changes. This would involve the integration of streaming data technologies and real-time machine learning algorithms.

Exploring the integration of other machine learning techniques, such as reinforcement learning and deep learning models, could also yield valuable insights. These advanced techniques might improve the models' ability to learn from complex, high-dimensional data and make more accurate predictions.

Finally, conducting longitudinal studies to track the performance of sentiment-based portfolios over longer periods and across different market cycles would provide a deeper understanding of the long-term viability and effectiveness of these approaches. This could also help to identify any potential limitations or biases in the models and suggest areas for further refinement.

In conclusion, future work should aim to broaden the scope of analysis, enhance data quality and processing capabilities, incorporate weighted sentiment analysis, and explore advanced machine learning techniques. By addressing these areas, researchers can continue to unlock the potential of machine learning in providing valuable insights into the dynamic and evolving cryptocurrency market.

Bibliography

Abdaljalil, S. and Bouamor, H. (2021). An exploration of automatic text summarization of financial reports, *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pp. 1–7.

Abraham, J., Higdon, D., Nelson, J. and Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis, *SMU Data Science Review* **1**(3): 1.

Adhikari, S., Thapa, S., Naseem, U., Lu, H. Y., Bharathy, G. and Prasad, M. (2023). Explainable hybrid word representations for sentiment analysis of financial news, *Neural Networks* **164**: 115–123.

Agarwal, A., Vats, S., Agarwal, R., Ratra, A., Sharma, V. and Gopal, L. (2023). Sentiment analysis in stock price prediction: a comparative study of algorithms, *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 1403–1407.

Alslaity, A. and Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions, *Behaviour & Information Technology* **43**(1): 139–164.

Alvarez, R., Garcia, D., Moreno, Y. and Schweitzer, F. (2015). Sentiment cascades in the 15m movement, *EPJ Data Science* **4**: 1–13.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models, *arXiv preprint arXiv:1908.10063* .

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, *2024 Third International Conference on Distributed Computing and High Performance Computing (DCHPC)*, IEEE, pp. 1–7.

Aziz, S., Dowling, M., Hammami, H. and Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach, *European Financial Management* **28**(3): 744–770.

Bahja, M. (2020). Natural language processing applications in business, *E-Business-Higher Education and Intelligence Applications* .

Biju, A., Mathew, A. M., Nithi Krishna, P. and Akhil, M. (2022). Is the future of bitcoin safe? a triangulation approach in the reality of btc market through a sentiments analysis, *Digital Finance* pp. 1–16.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market, *Journal of computational science* **2**(1): 1–8.

Bouri, E., Molnár, P., Azzi, G., Roubaud, D. and Hagfors, L. I. (2017). On the hedge and safe haven properties of bitcoin: Is it really more than a diversifier?, *Finance Research Letters* **20**: 192–198.

Bozanta, A., Angco, S., Cevik, M. and Basar, A. (2021a). Sentiment analysis of stock-twits using transformer models, *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 1253–1258.

Bozanta, A., Angco, S., Cevik, M. and Basar, A. (2021b). Sentiment analysis of stock-twits using transformer models, *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 1253–1258.

Buterin, V. et al. (2014). A next-generation smart contract and decentralized application platform, *white paper* **3**(37): 2–1.

Chahooki, M. A. Z. and KJ, T. S. (2023). Cryptocurrencies investment framework using sentiment analysis of twitter influencers, *Indonesian Journal of Electrical Engineering and Computer Science* **30**(2): 7.

Chen, Y.-J., Wu, C.-H., Chen, Y.-M., Li, H.-Y. and Chen, H.-K. (2017). Enhancement of fraud detection for narratives in annual reports, *International Journal of Accounting Information Systems* **26**: 32–45.

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

Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, *Proceedings of the 2008 international conference on web search and data mining*, pp. 231–240.

Dodevska, L., Petreski, V., Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. and Trajanov, D. (2019). Predicting companies stock price direction by using sentiment analysis of news articles, *Proceedings of the 15th Annual International Conference on Computer Science and Education in Computer Science*, pp. 37–42.

Du, X. and Tanaka-Ishii, K. (2020). Stock embeddings acquired from news articles and price history, and an application to portfolio optimization, *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3353–3363.

Guo, H., Zhang, D., Liu, S., Wang, L. and Ding, Y. (2021). Bitcoin price forecasting: A perspective of underlying blockchain transactions, *Decision Support Systems* **151**: 113650.

Guo, X. and Li, J. (2019). A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency, *2019 Sixth International*

Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, pp. 472–477.

Gurrib, I. and Kamalov, F. (2022). Predicting bitcoin price movements using sentiment analysis: a machine learning approach, *Studies in Economics and Finance* **39**(3): 347–364.

Hartmann, J., Heitmann, M., Siebert, C. and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis, *International Journal of Research in Marketing* **40**(1): 75–87.

Hazourli, A. (2022). Financialbert-a pretrained language model for financial text mining, *Technical report*, Technical report.

Huang, A. H., Wang, H. and Yang, Y. (2022). Finbert: A large language model for extracting information from financial text, *Contemporary Accounting Research* .

Huang, A. H., Wang, H. and Yang, Y. (2023). Finbert: A large language model for extracting information from financial text, *Contemporary Accounting Research* **40**(2): 806–841.

Huang, J.-Y. and Liu, J.-H. (2020). Using social media mining technology to improve stock price forecast accuracy, *Journal of Forecasting* **39**(1): 104–116.

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z. and Zhang, J. (2021). Lstm based sentiment analysis for cryptocurrency prediction, *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26*, Springer, pp. 617–621.

Hung, L. P. and Alias, S. (2023). Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection, *Journal of Advanced Computational Intelligence and Intelligent Informatics* **27**(1): 84–95.

Inamdar, A., Bhagtani, A., Bhatt, S. and Shetty, P. M. (2019). Predicting cryptocurrency value using sentiment analysis, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, pp. 932–934.

Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M. and Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach, *Artificial Intelligence Review* **57**(3): 62.

Jin, Z., Yang, Y. and Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and lstm, *Neural Computing and Applications* **32**: 9713–9729.

Jung, H. S., Lee, S. H., Lee, H. and Kim, J. H. (2023). Predicting bitcoin trends through machine learning using sentiment analysis with technical indicators., *Computer Systems Science & Engineering* **46**(2).

Karalevicius, V., Degrande, N. and De Weerdt, J. (2018). Using sentiment analysis to predict interday bitcoin price movements, *The Journal of Risk Finance* **19**(1): 56–75.

Koutmos, D. (2018). Bitcoin returns and transaction activity, *Economics Letters* **167**: 81–85.

Kurani, A., Doshi, P., Vakharia, A. and Shah, M. (2023). A comprehensive comparative study of artificial neural network (ann) and support vector machines (svm) on stock forecasting, *Annals of Data Science* **10**(1): 183–208.

Lee, H. S. (2020). Exploring the initial impact of covid-19 sentiment on us stock market using big data, *Sustainability* **12**(16): 6648.

Li, D., Wang, Y., Madden, A., Ding, Y., Tang, J., Sun, G. G., Zhang, N. and Zhou, E. (2019). Analyzing stock market trends using social media user moods and social influence, *Journal of the Association for Information Science and Technology* **70**(9): 1000–1013.

Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014). News impact on stock price return via sentiment analysis, *Knowledge-Based Systems* **69**: 14–23.

Liu, S., Li, T., Li, Z., Srikumar, V., Pascucci, V. and Bremer, P.-T. (2018). Visual interrogation of attention-based models for natural language inference and machine comprehension, *Technical report*, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

Liu, Y. and Tsyvinski, A. (2021). Risks and returns of cryptocurrency, *The Review of Financial Studies* **34**(6): 2689–2727.

Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L.-P. and Benoit, D. F. (2021). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data, *Machine Learning* pp. 1–24.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts, *Journal of the Association for Information Science and Technology* **65**(4): 782–796.

McMillan, B., Myers, J., Nguyen, A., Robinson, D. and Kennard, M. (2022). Analysis and comparison of natural language processing algorithms as applied to bitcoin conversations on social media, *The Journal of Investing* **31**(2): 38–59.

Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams engineering journal* **5**(4): 1093–1113.

Mehta, P., Pandya, S. and Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning, *PeerJ Computer Science* **7**: e476.

Mittal, A. and Goel, A. (2012). Stock prediction using twitter sentiment analysis, *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* **15**: 2352.

Naeem, M. A., Mbarki, I. and Shahzad, S. J. H. (2021). Predictive role of online investor sentiment for cryptocurrency market: Evidence from happiness and fears, *International Review of Economics & Finance* **73**: 496–514.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system, *Decentralized business review* p. 21260.

Nofer, M., Gomber, P., Hinz, O. and Schiereck, D. (2017). Blockchain, *Business & Information Systems Engineering* **59**: 183–187.

Oliveira, N., Cortez, P. and Areal, N. (2013). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter, *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pp. 1–8.

Oliveira, N., Cortez, P. and Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures, *Decision Support Systems* **85**: 62–73.

Omuya, E. O., Okeyo, G. and Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing, *Engineering Reports* **5**(3): e12579.

Pano, T. and Kashef, R. (2020). A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19, *Big Data and Cognitive Computing* **4**(4): 33.

Parekh, R., Patel, N. P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I. E. and Sharma, R. (2022). Dl-guess: Deep learning and sentiment analysis-based cryptocurrency price prediction, *IEEE Access* **10**: 35398–35409.

Parveen, S., Satti, Z. W., Subhan, Q. A. and Jamil, S. (2020). Exploring market over-reaction, investors' sentiments and investment decisions in an emerging stock market, *Borsa Istanbul Review* **20**(3): 224–235.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion, *Information fusion* **37**: 98–125.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z. and Wu, Q. J. (2022). A review of generalized zero-shot learning methods, *IEEE transactions on pattern analysis and machine intelligence* **45**(4): 4051–4070.

Prajapati, P. (2020). Predictive analysis of bitcoin price considering social sentiments, *arXiv preprint arXiv:2001.10343* .

Pérez, J. M., Giudici, J. C. and Luque, F. (2021). pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Raju, S. and Tarif, A. M. (2020a). Real-time prediction of bitcoin price using machine learning techniques and public sentiment analysis, *arXiv preprint arXiv:2006.14473* .

Raju, S. and Tarif, A. M. (2020b). Real-time prediction of bitcoin price using machine learning techniques and public sentiment analysis, *arXiv preprint arXiv:2006.14473* .

Ren, R. and Wu, D. (2018). An innovative sentiment analysis to measure herd behavior, *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **50**(10): 3841–3851.

Ren, R., Wu, D. D. and Liu, T. (2018a). Forecasting stock market movement direction using sentiment analysis and support vector machine, *IEEE Systems Journal* **13**(1): 760–770.

Ren, R., Wu, D. D. and Liu, T. (2018b). Forecasting stock market movement direction using sentiment analysis and support vector machine, *IEEE Systems Journal* **13**(1): 760–770.

Rupande, L., Muguto, H. T. and Muzindutsi, P.-F. (2019). Investor sentiment and stock return volatility: Evidence from the johannesburg stock exchange, *Cogent Economics & Finance* **7**(1): 1600233.

Sattarov, O., Jeon, H. S., Oh, R. and Lee, J. D. (2020). Forecasting bitcoin price fluctuation by twitter sentiment analysis, *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, IEEE, pp. 1–4.

Serrano-Cinca, C., Gutiérrez-Nieto, B. and López-Palacios, L. (2015). Determinants of default in p2p lending, *PloS one* **10**(10): e0139427.

Shuhidan, S. M., Hamidi, S. R., Kazemian, S., Shuhidan, S. M. and Ismail, M. A. (2018). Sentiment analysis for financial news headlines using machine learning algorithm, *Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018: KEER 2018, 19-22 March 2018, Kuching, Sarawak, Malaysia*, Springer, pp. 64–72.

Souma, W., Vodenska, I. and Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods, *Journal of Computational Social Science* **2**(1): 33–46.

Subbaiah, B., Murugesan, K., Saravanan, P. and Marudhamuthu, K. (2024). An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network, *Artificial Intelligence Review* **57**(2): 34.

Swan, M. (2015). *Blockchain: Blueprint for a new economy*, " O'Reilly Media, Inc.".

Tasca, P., Hayes, A. and Liu, S. (2018). The evolution of the bitcoin economy: Extracting and analyzing the network of payment relationships, *The Journal of Risk Finance* **19**(2): 94–126.

Tetlock, P. C., Saar-Tsechansky, M. and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals, *The journal of finance* **63**(3): 1437–1467.

Tumarkin, R. and Whitelaw, R. F. (2001). News or noise? internet postings and stock prices, *Financial Analysts Journal* **57**(3): 41–51.

Vakil, T. L. (2019). *Can Disaggregation in the Financial Statements Enhance the Credibility and Quality of Non-Gaap Disclosures?*, PhD thesis.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.

Vo, A.-D., Nguyen, Q.-P. and Ock, C.-Y. (2019). Sentiment analysis of news for effective cryptocurrency price prediction, *International Journal of Knowledge Engineering* **5**(2): 47–52.

Wang, C., Han, D., Liu, Q. and Luo, S. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm, *IEEE Access* **7**: 2161–2168.

Wang, Y., Yao, Q., Kwok, J. T. and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning, *ACM computing surveys (csur)* **53**(3): 1–34.

Wu, S., Liu, Y., Zou, Z. and Weng, T.-H. (2022). S_i_lstm: stock price prediction based on multiple data sources and sentiment analysis, *Connection Science* **34**(1): 44–62.

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z. and Wang, H. (2023). Empirical study of zero-shot ner with chatgpt, *arXiv preprint arXiv:2310.10035* .

Yao, W., Xu, K. and Li, Q. (2019). Exploring the influence of news articles on bitcoin price with machine learning, *2019 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, pp. 1–6.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M. and Liu, X.-Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models, *Proceedings of the fourth ACM international conference on AI in finance*, pp. 349–356.

Zhang, H., Wang, X., Liu, J., Zhang, L. and Ji, L. (2023). Chinese named entity recognition method for the finance domain based on enhanced features and pretrained language models, *Information Sciences* **625**: 385–400.

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.

Zhang, W. (2011). *News based forecasting and modeling*, PhD thesis, State University of New York at Stony Brook.

Zhao, Z., Hao, Z., Wang, G., Mao, D., Zhang, B., Zuo, M., Yen, J. and Tu, G. (2021). Sentiment analysis of review data using blockchain and lstm to improve regulation for a sustainable market, *Journal of Theoretical and Applied Electronic Commerce Research* **17**(1): 1–19.

Zhou, Z.-H. (2021). *Machine learning*, Springer Nature.