Technische Hochschule Deggendorf
Fakultät Angewandte Informatik

Studiengang Master of Science (M.Sc.) in Artificial Intelligence and Data Science

Auswirkung des sozialen Konsenses auf die Bitcoin-Preisaktion

Impact of Social Consensus on Bitcoin Price Action

Masterarbeit zur Erlangung des akademischen Grades:

Master of Science (M.Sc.)

an der Technischen Hochschule Deggendorf

Vorgelegt von:
Cihan Evren

Matrikelnummber:
00821461

Prüfungsleitung:
Glauner, Prof. Dr. Patrick

Ergänzende Prüfende:
Hable, Dr. Robert

Am: September 1, 2023

Declaration

TECHNISCHE
HOCHSCHULE
DEGGENDORF

Erklärung

Name des Studierenden: Cihan Evren

Name des Betreuenden: Prof. Dr. Patrick Glauner

Thema der Abschlussarbeit:

Auswirkung des sozialen Konsenses auf die Bitcoin-Preisaktion ....
....
....
....

1. Ich erkläre hiermit, dass ich die Abschlussarbeit gemäß § 35 Abs. 7 RaPO (Rahmenprüfungsordnung für die Fachhochschulen in Bayern, BayRS 2210-4-1-4-1-WFK) selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

   Deggendorf, 01.09.2023 ....
   Datum — Unterschrift des Studierenden

2. Ich bin damit einverstanden, dass die von mir angefertigte Abschlussarbeit über die Bibliothek der Hochschule einer breiteren Öffentlichkeit zugänglich gemacht wird:

   ○ Nein

   ☒ Ja, nach Abschluss des Prüfungsverfahrens

   ○ Ja, nach Ablauf einer Sperrfrist von ... Jahren.

   Deggendorf, 01.09.2023 ....
   Datum — Unterschrift des Studierenden

Bei Einverständnis des Verfassenden vom Betreuenden auszufüllen:

Eine Aufnahme eines Exemplars der Abschlussarbeit in den Bestand der Bibliothek und die Ausleihe des Exemplars wird:

○ Befürwortet

○ Nicht befürwortet

Deggendorf, ...
Datum — Unterschrift des Betreuenden

Acknowledgements

I would like to express my gratitude to all my professors at Technische Hochschule Deggendorf and the University of South Bohemia. Their invaluable guidance and teachings have helped me expand my skillset in Artificial Intelligence and Data Science.

I would especially like to thank my supervisor, Professor Patrick Glauner, for believing in my thesis proposal and providing unwavering guidance and support throughout this journey.

Last but not least, I wish to extend my heartfelt appreciation to my parents, my brother, and my friends. Their constant support, encouragement, and unique perspectives have been instrumental in broadening my horizons and ensuring the success of this research.

Abstract

This research delves into the untapped potential of YouTube comments as a novel predictive metric for Bitcoin price fluctuations. By sourcing comments from influential cryptocurrency-related YouTube channels spanning from 2019 to 2023, the study offers a comprehensive analysis of public sentiment and its correlation with Bitcoin's market behavior. To achieve this, the research employs state-of-the-art AI models, notably RoBERTa for binary classification and TimeLM for sentiment analysis. The methodology is meticulously designed, beginning with data preprocessing to filter relevant comments, followed by sentiment categorization, and culminating in a correlation analysis with Bitcoin's price trends. Preliminary results are promising, indicating a discernible relationship between the sentiment derived from YouTube comments and Bitcoin price movements. The Granger causality analysis further bolsters these findings, suggesting that historical sentiment data can be instrumental in forecasting Bitcoin's future price trajectory. However, it is paramount to acknowledge that Bitcoin's price dynamics are multifaceted, influenced by a myriad of factors ranging from supply-demand dynamics to regulatory changes. Yet, this research underscores the pivotal role of social media sentiment, particularly from platforms like YouTube, in the intricate web of cryptocurrency market predictors. In conclusion, this study not only highlights the significance of YouTube sentiment as a valuable metric in the predictive matrix but also paves the way for future research in leveraging social media data for financial market analysis.

Keywords: Bitcoin, Sentiment Analysis, Artificial Intelligence, TimeLM, RoBERTa, YouTube

Contents

List of Figures

List of Tables

List of Listings

Chapter 1

Introduction

1.1 Background

In a paper released under the pseudonym Satoshi Nakamoto, Bitcoin was introduced to the global financial landscape [Nakamoto, 2009]. Since its official launch on January 3, 2009, Bitcoin's rise has led to many other cryptocurrencies being developed. By 2023, there are approximately 23,000 distinct cryptocurrencies cataloged by CoinMarketcap. This rapid expansion, driven largely by the market's inherent volatility, has captivated a vast array of individuals, many of whom are primarily motivated by potential profits. As the cryptocurrency realm expanded, enthusiasts began to congregate on various social media platforms to exchange news, insights, and opinions. Platforms like Twitter, Reddit, and YouTube emerged as primary hubs for these discussions.

However, predicting Bitcoin's price trajectory remains a complex endeavor. The price is swayed by a confluence of factors, from regulatory shifts and macroeconomic trends to technological innovations and market demand. Amidst this complexity, sentiment analysis has carved out a significant niche, serving as a lens through which the collective mood and perspectives of potential investors can be discerned. While Twitter and Reddit have traditionally been the go-to platforms for sentiment analysis, they capture only a segment of the broader sentiment spectrum. Recognizing this gap, this study delves into the relatively untapped domain of YouTube sentiment analysis.

YouTube, renowned as the world's premier video-sharing platform, houses a vast reservoir of sentiment data in the form of user comments. Yet, its potential as a sentiment source for cryptocurrency analysis remains largely overlooked. This research endeavors to harness this unexplored potential, conducting a thorough analysis of comments from leading YouTube channels centered on cryptocurrency over four years (2019-2023). Utilizing cutting-edge AI models, such as RoBERTa and TimeLM, the study distills and categorizes sentiments, presenting a novel perspective on public sentiment towards Bitcoin.

1.2 Motivation

The primary motivation behind this research is to augment the existing literature with an innovative metric for predicting Bitcoin's price. While traditional platforms like Twitter and Reddit have been extensively studied, the sentiment landscape remains incomplete without considering other influential platforms. YouTube, with its vast user base and dynamic content, offers a unique perspective that has been largely overlooked. By tapping into this reservoir of sentiment data, this study aims to provide a more holistic understanding of the factors influencing Bitcoin's price.

Harnessing advanced AI models, the research meticulously filters Bitcoin-related comments from YouTube, ensuring the data's relevance and accuracy. Each comment is then assigned a sentiment score, translating the vast array of opinions into quantifiable data. This process not only captures the prevailing sentiment towards Bitcoin on YouTube but also allows for a nuanced analysis of how this sentiment correlates with Bitcoin's price movements. By comparing the sentiment scores with Bitcoin's price action, the study seeks to uncover patterns and relationships that might have otherwise remained obscured.

To further solidify the findings, a lead-lag analysis is conducted. This rigorous statistical approach confirms the temporal relationship between YouTube sentiment and Bitcoin's price fluctuations. By determining which of the two leads or lags, the study provides invaluable insights into the predictive power of YouTube sentiment. In essence, this research not only introduces YouTube as a pivotal sentiment source but also underscores its potential to enhance the accuracy and depth of Bitcoin price predictions.

1.3 Workflow

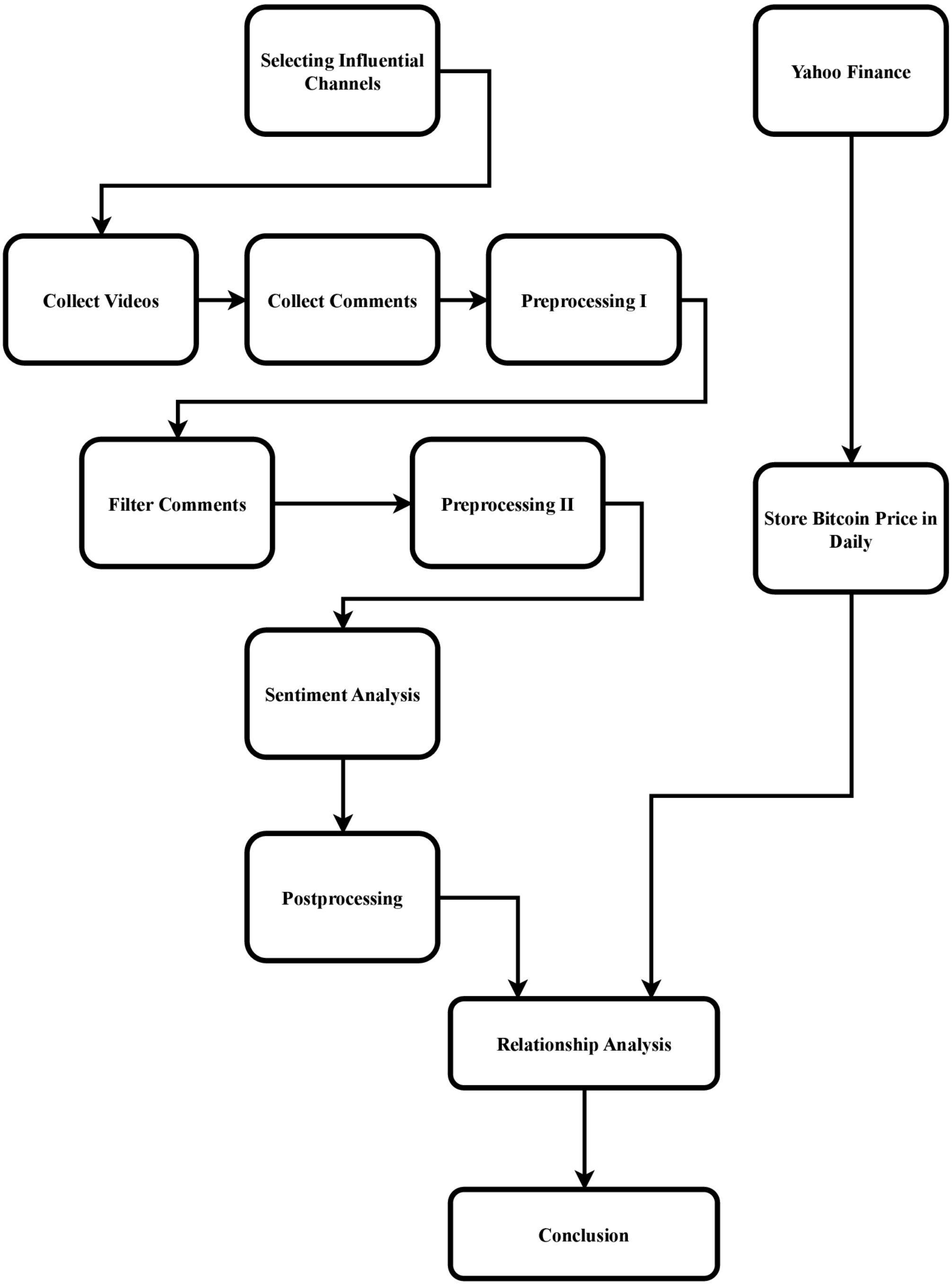


Figure 1.1: Workflow

Chapter 2

Literature Review

The study "Does Sentiment Impact Cryptocurrency" by Anamika, Madhumita Chakraborty, and Sowmya Subramaniam challenges the traditional economic theory that investors are rational and that markets are efficient [Anamika et al., 2023]. They argue that investor psychology plays a significant role in asset markets, including the cryptocurrency market. The authors found that Bitcoin sentiment has a positive impact on Bitcoin returns, indicating that sentiment plays a significant role in determining its prices. The sentiment of Bitcoin significantly influences the returns of other cryptocurrencies like Ethereum, Ripple, Litecoin, and Bitcoin-Cash. The study concludes that investor sentiment, particularly towards Bitcoin, has a significant impact on cryptocurrency prices. It also found that sentiment in the equity market can influence cryptocurrency prices, suggesting that cryptocurrencies can serve as an alternative investment avenue during bearish market conditions.

Building on the idea of sentiment influencing Bitcoin prices, another study titled "A Stacking Ensemble Deep Learning Model for Bitcoin Price Prediction Using Twitter Comments on Bitcoin" introduces a novel ensemble deep learning model for predicting Bitcoin's price [Ye et al., 2022]. The model uses price data, technical indicators, and sentiment indexes derived from social media texts. The authors found that the ensemble method has better performance and can better assist investors in making the right investment decision than other traditional models. They conclude that their model provides a more accurate prediction of Bitcoin's price.

In the same vein, the research titled "Attention Transformer with Sentiment on Cryptocurrencies Price Prediction" investigates the forecasting capability of the Transformer model on Bitcoin (BTC) and Ethereum (ETH) price data [Zhao. et al., 2022]. The authors compared the performance of the Transformer model with the Long Short Term Memory (LSTM) model. They also explored the impact of sentiment analysis on the model's performance in forecasting future prices. The sentiment scores are derived from Twitter data using the Valence Aware Dictionary and sEntiment Reasoner (VADER). The authors concluded that while the LSTM model outperformed the proposed Transformer model on predicting BTC and ETH future prices, sentiment analysis could improve the Transformer model's performance on BTC price prediction.

In the 2021 research by Akyildirim and his colleagues, the interconnectedness between cryptocurrency returns and investor sentiments is meticulously examined [Akyildirim et al., 2021]. The study employs the novel cryptocurrency-specific MarketPsych sentiment data for 13 major cryptocurrencies, analyzing their dynamic network connectedness. Key findings reveal that cryptocurrencies with higher market capitalization play a dominant role in the overall directional connectedness. Interestingly, while Bitcoin has traditionally been the dominant cryptocurrency, the study indicates a shift in its dominance in return spillovers, with alt-coins gaining prominence. However, Bitcoin remains the primary transmitter of sentiment shocks. The research underscores the significance of sentiments in the cryptocurrency market, suggesting that investor sentiments play a pivotal role in influencing cryptocurrency prices.

The article titled "Bitcoin Price Change and Trend Prediction Through Twitter Sentiment and Data Volume" by Jacques Vella Critien, Albert Gatt, and Joshua Ellul explores the potential of Twitter sentiment and data volume in predicting Bitcoin price changes and trends [Vella Critien et al., 2022]. The authors found that their model can predict not only the direction of the price change but also the magnitude of the price change with relative accuracy (63%). They also found that this model yields more reliable predictions when used alongside a price trend prediction model. The authors conclude that Twitter sentiment and data volume can be useful tools in predicting Bitcoin price changes and trends.

Another paper that mainly focuses on collective emotions explores the influence of investors' emotions and sentiment on Bitcoin volatility and the characterization of Bitcoin bubble phases [Bourghelle et al., 2022]. The study found that collective emotions play a crucial role in the formation and collapse of the Bitcoin bubble. The study concluded that the power of sentiment has a time-varying effect on the market. In a calm state, where Bitcoin volatility is relatively low, collective emotions have a negative impact on Bitcoin volatility, prompting stability. However, during a bubble formation, the effect of emotions turns significantly positive as investors become less fearful and more reassured, which can increase volatility and destabilize the market.

Balcilar, Bouri, Gupta, and Roubaud, in their 2017 research, investigate the potential of trading volume as a predictor for Bitcoin returns and volatility [Balcilar et al., 2017]. Unlike previous studies that focused on the conditional mean of the returns distribution, this research employs a non-parametric causality-in-quantiles test to analyze the relationship across their entire respective conditional distributions. Spanning data from 19th December 2011 to 25th April 2016, the study uncovers that trading volume can predict Bitcoin returns, but not consistently across all market regimes. Specifically, volume predictions falter in both bear and bull market conditions. Furthermore, the research indicates that trading volume does not predict the volatility of Bitcoin returns at any point in the conditional distribution. This study underscores the significance of considering nonlinearity and tail behavior when examining causal relationships between Bitcoin returns and trading volume.

In Niranjan Sapkota's 2022 research [Sapkota, 2022], the influence of news media sentiments on Bitcoin volatility is meticulously examined. The study leverages three distinct range-based volatility estimates and two sentiment types, namely psychological and financial sentiments, utilizing four sentiment dictionaries. Analyzing 17,490 news articles from 91 major English-language newspapers listed in the LexisNexis database spanning from January 2012 to August 2021, Sapkota identifies a significant correlation between news media sentiments and Bitcoin volatility. Utilizing the heterogeneous autoregressive model for realized volatility (HAR-RV) with news sentiments as explanatory variables, the study demonstrates enhanced model fit and forecasting precision. Notably, psychological sentiments exhibit medium-term effects, while financial sentiments have long-term implications on Bitcoin volatility. The National Research Council Emotion Lexicon pinpoints anticipation and trust as the primary emotional drivers influencing Bitcoin volatility.

A study conducted by [Mokni et al., 2022] applies a quantile-based analysis to investigate the causal relationships between Bitcoin and investor sentiment. The study concludes that the COVID-19 crisis has no effect on the causal relationship between Bitcoin and investor sentiment. However, Bitcoin returns/volatility have significant predictive power on the investor sentiment. The study also shows that investor sentiment and Bitcoin prices are significantly affected by the COVID-19 pandemic.

In the 2023 study by Jiang, Rodríguez Jr., and Zhang, the intricate relationship between macroeconomic fundamentals, equity markets, and cryptocurrency prices is explored [Jiang et al., 2023]. The research is grounded in asset pricing theory, suggesting that equity and cryptocurrency markets share a common fundamental. Through cointegration tests, the study reveals that consumption, a pivotal asset pricing primitive, serves as this shared fundamental. The research further identifies that deviations from these fundamentals are influenced not just by speculative sentiment but also by macroeconomic factors and time-varying uncertainty. Contrary to the prevailing belief that cryptocurrency prices are solely driven by speculative behavior, this study establishes a long-run equilibrium relationship between equity prices, cryptocurrency prices, and macroeconomic fundamentals. The authors propose three channels linking equity markets, cryptocurrency markets, and the macroeconomy: portfolio allocation decisions, intermarket order flows, and technological adaptation expectations.

The article ”Sentiment, Google queries and explosivity in the cryptocurrency market” discusses the speculative bubble phenomena in the cryptocurrency market, which is often associated with alternating phases of investors’ fear and greed [Agosto et al., 2022]. The authors propose using information derived from a large set of cryptocurrency news and Google Search Indices to detect and possibly anticipate the presence of speculative bubbles in cryptocurrency prices. The results show that sentiment can predict speculative bubble occurrences.

The paper ”On the predictive power of tweet sentiments and attention on Bitcoin” investigates the predictive power of information contained in social media tweets on Bitcoin market dynamics [Suardi et al., 2022]. The study finds a significant positive relationship between the conditional volatility of bitcoin returns and sentiment dispersion. This finding indicates that Bitcoin investors face greater risk when exposed to higher uncertainty levels. The refined investor attention measure has only significant predictive power on Bitcoin trading volume. The study concludes that using investor attention alone as a trading parameter does not produce superior performance over the long term.

In Oikonomopoulos Sotirios’ 2022 study on ”Cryptocurrency price prediction using social media sentiment analysis,” the rapid growth of cryptocurrencies and the influential role of Twitter sentiment are explored [Oikonomopoulos et al., 2022]. The research, centered on seven major cryptocurrencies, employed the Valence Aware Dictionary for Sentiment Reasoning (VADER) to analyze tweet sentiments. The findings highlighted a distinct relationship between sentiment and price fluctuations. While sentiment followed price changes for Bitcoin, Cardano, XRP, and Doge, it predicted price movements for Ethereum and Polkadot, with the latter two achieving prediction accuracies of 99.67% and 99.17% respectively.

In Junwei Chen’s 2023 research [Chen, 2023], the intricate relationship between Bitcoin’s price and various influencing factors is explored using machine learning techniques. The study’s primary objective is to derive a predictive model with high accuracy for forecasting Bitcoin’s price on the subsequent day. To achieve this, Chen employs two prominent algorithms: random forest regression and LSTM. The research underscores the significance of understanding which variables predominantly influence Bitcoin's price. While prior studies have leaned towards the ARMA model of time series and deep learning's LSTM algorithm, Chen's findings suggest that random forest regression offers superior prediction errors in terms of RMSE and MAPE compared to LSTM. Notably, the research identifies a shift in influential variables over time. From 2015 to 2018, Bitcoin's price was influenced by major US stock market indexes and the price of ETH. Post-2018, the determining variables transitioned to ETH price and the Japanese stock market index JP225.

In conclusion, the studies collectively underscore the significant influence of investor sentiment on Bitcoin prices and volatility. They highlight the potential of sentiment analysis, derived from various sources such as Twitter, Google Search Indices, and investor surveys, for predicting Bitcoin prices and trends. However, the effectiveness of these predictive models can vary based on the specific sentiment measures used, the data sources, and the prevailing market conditions. Notably, none of these studies have utilized YouTube as a data source for sentiment analysis, indicating a gap in the current literature. My primary focus is to fill this gap by exploring the potential of YouTube data in predicting Bitcoin prices and trends, thereby contributing a novel perspective to the existing body of knowledge.

Chapter 3

Data Collection And Preprocessing

Data collection is an integral part of sentiment analysis, serving as the foundation upon which the entire research process is built. The raw data collected forms the basis for subsequent stages of the research, including data cleaning, analysis, and interpretation. The quality, relevance, and comprehensiveness of this data can significantly influence the validity and reliability of the research findings. In this study, the main focus is concentrated on the collection of two primary types of data: social media sentiment data and financial data.

Within the scope of sentiment data, the focus is to utilize YouTube, a platform that has often been overlooked in academic research despite its extensive user base and popularity. As one of the world's most frequented online platforms, YouTube boasts a monthly logged-in user base of over two billion individuals. The platform's active comment sections serve as a vibrant forum for public discourse, providing a rich and diverse source of sentiment data. By mining and analyzing the sentiments expressed in these comments, the intention is to tap into the collective mood and opinions of a large, globally distributed population.

On the financial data front, the Python-based yfinance library is leveraged. This robust tool facilitates the efficient retrieval of historical market data from Yahoo Finance, a vital resource for the study's objective to investigate the correlation between public sentiment and Bitcoin prices. To this end, Bitcoin's historical data, specifically daily opening prices and corresponding dates, is meticulously extracted over a span of four years. This concise selection of data offers an insightful perspective on Bitcoin's price trajectory and market trends during this defined period, aligning closely with the study's focused exploration of financial patterns.

3.1 Social Media Sentiment Data

The social media sentiment data is collected with the aim of running a sentiment analysis to understand its effect on Bitcoin's price formation. Social media platforms are a treasure trove of public sentiment, with users freely expressing their opinions, reactions, and predictions about various topics, including cryptocurrencies. By analyzing these sentiments, significant insights can be gained into the collective mood of the market and how it might influence Bitcoin's price.

3.1.1 Determining The Source

While many studies have focused on sentiment analysis using data from platforms like Reddit, Google Trends, or Twitter, YouTube comments have been largely overlooked. This is despite the fact that YouTube has a vast community of users who actively engage in discussions about a wide range of topics, including cryptocurrencies. One reason for this oversight may be the unique challenges posed by YouTube as a data source. Unlike Twitter, where hashtags make it relatively easy to collect topic-specific data, YouTube comments do not typically use hashtags. This lack of structured tagging makes it more challenging to collect topic-specific comments. However, I believe that these challenges can be overcome with careful data collection strategies and sophisticated text analysis techniques. Therefore, in this study, YouTube comments have been selected as the primary source of social media sentiment data. The aim is to demonstrate that, despite its challenges, YouTube comments can provide valuable insights into public sentiment, contributing to a more comprehensive understanding of the factors influencing Bitcoin prices.

3.1.2 Identifying The Most Potent Channels

To ensure the quality and relevance of the data collected, eleven influential YouTube channels that focus on cryptocurrencies were identified. These channels were chosen based on their high numbers of subscribers and views, indicating their influence in the cryptocurrency community. The channels are listed in Table 3.1.

Number	Channel Name	Subscribers
1	BENJAMIN COWEN	784K
2	COIN BUREAU	2.29M
3	BITBOY CRYPTO	1.45M
4	CRYPTO LOVE	238K
5	LARK DAVIS	480K
6	CRYPTOSRUS	666K
7	CRYPTO BANTER	641K
8	ALTCOIN DAILY	1.31M
9	CRYPTO ZOMBIE	252K
10	CRYPTO FACE	216K
11	SCOTT MELKER	128K

Table 3.1: List of YouTube Channels with Subscriber Count

3.1.3 Scraping YouTube using YouTube API

The YouTube API plays a crucial role in this research, as it is the primary tool used to collect comments from videos published by various channels. These comments serve as a rich source of data for the sentiment analysis, providing insights into viewer reactions and opinions.

However, it is important to note that the YouTube API has certain limitations that affect the scope of the data collection part. One significant restriction is the inability to retrieve the historical view counts of a channel without explicit permission from the channel owner. This permission is granted in the form of an access key. Historical view counts could provide additional context for the study, as they offer a measure of the channel's popularity and viewer engagement over time.

The following part of the study will highlight how to use it efficiently to gather and make use of this data. The research relies on the YouTube API to collect the data needed, which includes pulling comments from numerous videos across various channels. The API, despite having some limitations, is still a strong tool for data collection.

The primary objective of the data collection process is to gather user comments posted under Bitcoin and cryptocurrency related YouTube videos. These comments serve as a valuable source of data for the sentiment analysis. These comments are collected daily, and while they may be posted at different times throughout the day, for the purpose of the study, all comments made within a single day are considered as belonging to that day. The data collection period spans from January 1, 2019, to January 1, 2023, providing a comprehensive four-year dataset.

To achieve this, the data collection process is divided into two distinct subtasks. The first subtask involves identifying the video IDs of the content published by each of the eleven channels under consideration. These video IDs serve as unique identifiers that allow for the location of each video on the YouTube platform. Over the four-year timespan from 2019 to 2023, a total of 15,341 videos from the eleven channels were identified. The distribution of videos across the channels can be seen in the Table 3.2 below:

Channel Name	**Number of Videos**
BENJAMIN COWEN	1812
COIN BUREAU	855
BITBOY CRYPTO	3445
CRYPTO LOVE	1160
LARK DAVIS	1566
CRYPTOSRUS	1475
CRYPTO BANTER	1418
ALTCOIN DAILY	1719
CRYPTO ZOMBIE	997
CRYPTO FACE	951
SCOTT MELKER	933

Table 3.2: Number of videos per channel from 2019 to 2023

The second subtask of the data collection process involves using the video IDs to access each individual video. Once a video has been accessed, the next step is to scrape all the user comments posted under it. This process is repeated for each video ID collected, allowing for the gathering of a comprehensive set of comments from each of the eleven channels.

It is important to note that the approach to data collection is designed to capture the dynamic nature of YouTube's comment section. First, all the comments for each video are collected, regardless of when the video was posted. Then, these comments are grouped by the day they were posted. This allows for capturing the total count of daily comments, even if they are posted under videos from different channels or under older videos. This approach acknowledges that users may visit and comment on older videos, and ensures that these comments are included in the dataset.

Over the four-year timespan of this study, a total of 2,839,528 comments were collected from 15,341 videos. The distribution of comments across the channels is as follows in Table 3.3:

Channel Name	**Number of Comments**
BENJAMIN COWEN	110634
COIN BUREAU	371788
BITBOY CRYPTO	678513
CRYPTO LOVE	87454
LARK DAVIS	315280
CRYPTOSRUS	273670
CRYPTO BANTER	218272
ALTCOIN DAILY	481380
CRYPTO ZOMBIE	326361
CRYPTO FACE	30696
SCOTT MELKER	36580

Table 3.3: Number of comments per channel from 2019 to 2023

The data we collected consists of several key elements: a unique identifier for each comment, the date the comment was made, the comment text itself, and the number of likes the comment received. A sample of this data is presented in Table 3.4 below:

ID	DATE	TEXT	LIKECOUNT
19326	24/10/2020	Let&#39;s go to 14k first before election	0
19336	24/10/2020	I would Truly prefer to Get back to atleast 12.4-.12.2 get that out the way and then shoot to the 14k plus area– but Truly who cares– Most of of us are Just Hodling.. Great TA as usual – Thx	5
19344	25/10/2020	you got 50k soon	1
19308	26/10/2020	I think when btc Hit the old ath from 2019 we make another 20% fomo pump to around 15 -16k and retest the 139k as support. Btc will leave this Chanel when we hit the 2019 ath. When the bears win and defeat the Bulls After double top we see the 12k and tray a retest in the incoming weeks.	0
19377	26/10/2020	Brilliant video as usual. Thank you really appreciate your content.	2
19380	26/10/2020	Great broad data analysis awesome content!	2
19383	26/10/2020	People are blinded by the fact LINK has a 1.38b total supply. It&#39;s already grossly overpriced. All it would take is for 500m to be released into circulating supply and it would lose over half its current value. No amount of TA can account for that risk.	2
19386	26/10/2020	Best Crypto analyst in the YouTube multiverse. Period. Fight me.	6
19394	26/10/2020	Always pristine content Love the risk metric of link/btc	3
19416	26/10/2020	Why do an analysis if there is Vectra Coin?	0

Table 3.4: Sample of Scraped Data

As can be observed from the sample data, the comments are in their raw form and have not been preprocessed yet. This means that the dataset includes elements such as emojis and HTML-encoded strings. For instance, the string `Let&#39;s` is an HTML-encoded representation of Let's, and emojis are represented by their respective Unicode characters.

More importantly, not all comments are directly related to Bitcoin or market conditions. Some comments are directed towards the video creator, praising the quality of the video or the analysis provided. For example, comments such as "Brilliant video as usual. Thank you really appreciate your content." and "Great broad data analysis awesome content!" are appreciative of the content creator rather than providing insights into Bitcoin or the market.

To utilize this dataset for the sentiment analysis, a number of preprocessing steps are required. Firstly, the data must be cleaned, which involves handling HTML-encoded strings and emojis. Secondly, the comments that are specifically related to Bitcoin and the broader cryptocurrency market need to be classified.

Given the sheer volume of comments - around 3 million - manually labelling each comment whether it is an opinion about Bitcoin or not would be a monumental task, both in terms of time and resources. To overcome this challenge, a binary classifier needs to be developed. This classifier will be trained to distinguish between comments that are relevant to our analysis (i.e., those that discuss Bitcoin and the cryptocurrency market) and those that are not (such as comments praising the video creator or discussing unrelated topics).

3.2 Financial Data

The financial data used in the study comprises the historical price data of Bitcoin, specifically the daily opening prices over a four-year period from January 1, 2019, to January 1, 2023. This data is sourced from Yahoo Finance, a reliable and widely used platform for financial information.

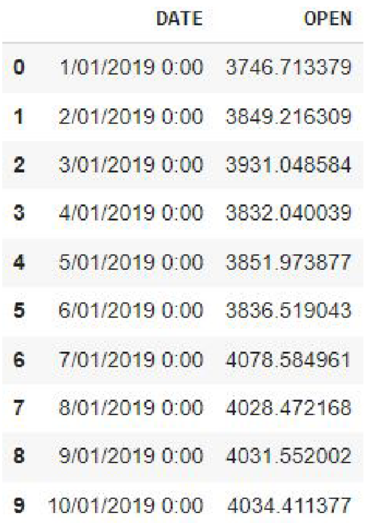
3.2.1 Financial Data Acquisition

To retrieve this data, the Python library yfinance [ranaroussi, 2023] is employed. This library provides a convenient interface to download historical market data from Yahoo Finance. Here is a Python code snippet demonstrating how to use yfinance to fetch the daily closing prices of Bitcoin in Listing 3.1 below:

Listing 3.1: Python Code to Fetch Financial Data

```python
import yfinance as yf

# Define the ticker symbol for Bitcoin on Yahoo Finance
ticker_symbol = 'BTC-USD'

# Define the period for which we want historical data
start_date = '2019-01-01'
end_date = '2023-01-01'

# Use yfinance to fetch the historical market data
btc = yf.Ticker(ticker_symbol)
hist = btc.history(start=start_date, end=end_date)

# Print the closing prices
print(hist['OPEN'])
```

In this script, the yfinance module is first imported and the ticker symbol for Bitcoin on Yahoo Finance, which is 'BTC-USD', is defined. The start and end dates for the period of interest are also specified. Then, the history method is called on the ticker object to fetch the historical market data for Bitcoin. Finally, the opening prices are printed as shown in Figure 3.1 below:



	DATE	OPEN
0	1/01/2019 0:00	3746.713379
1	2/01/2019 0:00	3849.216309
2	3/01/2019 0:00	3931.048584
3	4/01/2019 0:00	3832.040039
4	5/01/2019 0:00	3851.973877
5	6/01/2019 0:00	3836.519043
6	7/01/2019 0:00	4078.584961
7	8/01/2019 0:00	4028.472168
8	9/01/2019 0:00	4031.552002
9	10/01/2019 0:00	4034.411377

Figure 3.1: Bitcoin Historical Price

3.2.2 Financial Data Visualisation

Here, the historical price data of Bitcoin is explored, spanning a period of four years. This comprehensive dataset provides a detailed view of Bitcoin's valuation in USD over this the frame (Figure 3.2).

A noteworthy observation from this data is the two significant peaks in Bitcoin's price. The first peak occurred in May 2021, when Bitcoin reached a value of $64,000. This was followed by a second peak in November 2021, where Bitcoin's price soared to $69,000.

However, following these peaks, there has been a noticeable downward trend in Bitcoin's price. As of the most recent data point in December 2022, Bitcoin's price stands at $17,500. This marks a significant decrease from its previous highs in 2021.

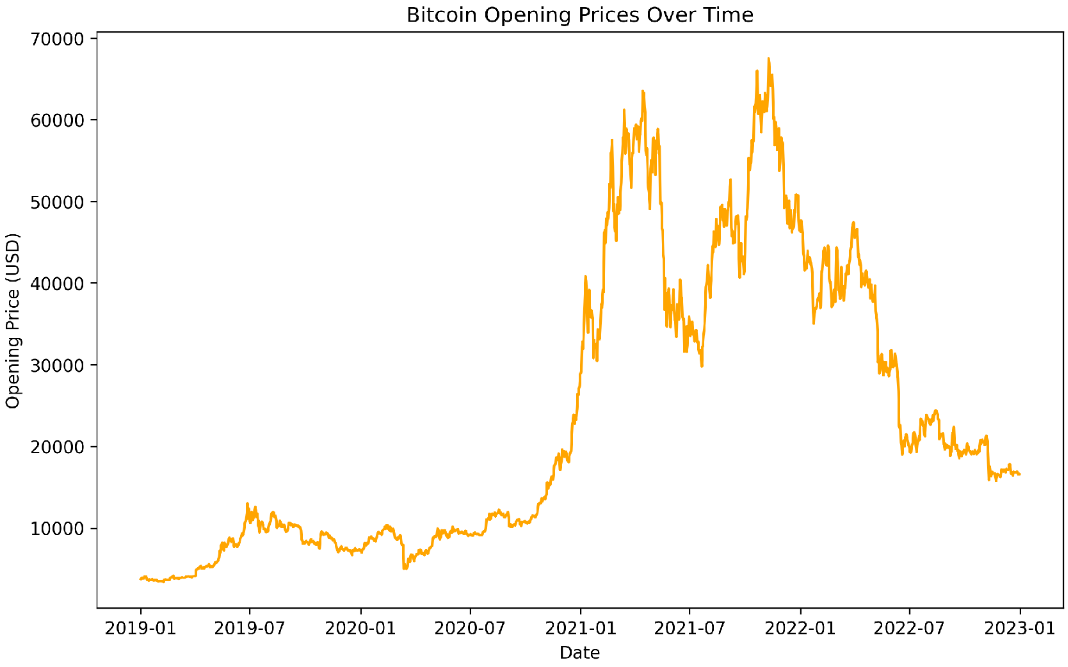


Figure 3.2: Bitcoin Historical Line Graph

3.3 Data Preprocessing Using PySpark

In the Data Preprocessing section, the essential procedures used to ready the dataset for sentiment analysis are detailed. The primary objective is to extract sentiments from the comments, a task that necessitates the resolution of two main challenges through the application of Artificial Intelligence (AI).

The first challenge involves the filtration of comments to retain only those pertinent to Bitcoin and the cryptocurrency market. This necessitates the elimination of comments that are predominantly about the content creator or the channel owner, as well as scam comments that attempt to deceive users into clicking on a misleading link. Given the sheer volume of our dataset, which comprises almost 3 million comments, manual labelling is impractical. Consequently, state-of-the-art AI models are used to construct a binary classifier capable of automating this process.

The second challenge is to ascertain the sentiment scores for each comment, after filtering the data to retain only comments about Bitcoin and the cryptocurrency market. This task can also be accomplished using AI.

Both of these challenges fall within the ambit of Natural Language Processing (NLP). To effectively train the mentioned models, it is imperative to preprocess the raw data and transform it into a format that the models can easily comprehend and learn from.

The preprocessing steps encompass replacing HTML entities, removing URLs, eliminating punctuation, replacing contractions, removing stop words, and lemmatizing words. These steps aid in cleaning and standardizing the data, rendering it more suitable for NLP tasks.

Given the extensive size of the dataset, the use of Python and pandas for data processing proved to be sluggish. Consequently, a distributed computing tool, PySpark [Foundation, 2023], which is better equipped to handle large datasets, was opted for.

In the subsequent sections, a code snippet that illustrates the functions used to implement these preprocessing steps will be shared. This will provide a more lucid understanding of how the raw data was transformed into a format conducive to training the AI models.

Furthermore, it is worth noting that the process of data preprocessing is not a one-size-fits-all approach. It requires careful consideration of the specific requirements of the task at hand and the characteristics of the data. In this case, the nature of the comments and the need to extract meaningful sentiments from them necessitated a series of preprocessing steps that might differ from those required for other NLP tasks.

Moreover, the choice of PySpark as the data processing tool underscores the importance of scalability in data preprocessing. As datasets grow larger, traditional data processing tools like Python and pandas may not be able to handle the increased volume efficiently. Distributed computing tools like PySpark, on the other hand, are designed to process large datasets, making them a more suitable choice for the task.

3.3.1 Application of Preprocessing Using PySpark

In this section, the presentation and discussion of code snippets for the preprocessing functions implemented using PySpark will be covered. PySpark, a Python library for Apache Spark, is a powerful tool for handling big data and performing distributed computing tasks. It is particularly useful for this research due to the large volume of comments being dealt with.

Data preprocessing is a critical step in any NLP task. It involves transforming raw text data into a format that can be easily understood and processed by AI models. The preprocessing steps can vary depending on the specific requirements of the task and the nature of the data. However, common steps often include removing unnecessary elements (like HTML tags and URLs), normalizing the text (such as converting all text to lower case and removing punctuation), and reducing words to their base or root form through a process called lemmatization.

Several User-Defined Functions (UDFs) were defined in PySpark to carry out these preprocessing steps. Each function is designed to perform a specific task, and when used together, they form a comprehensive preprocessing pipeline that prepares the raw comment data for subsequent analysis.

The UDFs below are part of the preprocessing steps that are taken to transform the raw data into preprocessed data.

Listing 3.2: Replace HTML Entities

```python
# Define UDFs
@udf(StringType())
def replace_html_entities(text):
  if text is not None:
    html_entities = {
        '&amp;': '&',
        '&#39;': "'",
        '&quot;': '"',
        '&lt;': '<',
    }
    pattern = re.compile('|'.join(html_entities.keys()))
    return pattern.sub(lambda x: html_entities[x.group()], text)
  else:
    return None
```

The function replaces HTML entities in the text with their corresponding characters. For example, it replaces ‘&amp;‘ with ‘&‘, ‘&#39;‘ with ‘’‘, and so on.

Listing 3.3: Remove URLs

```python
@udf(StringType())
def remove_url(text):
    if text is not None:
      return re.sub(r"http\S+", "", text)
    else:
      return None
```

The function removes URLs from the text. It uses a regular expression to match any string that starts with ‘http‘ and removes it.

Listing 3.4: Remove Punctuations

```
@udf(StringType())
def remove_punctuations(text):
    if text is not None:
      expression = [i.lower() for i in text if i not in punctuations]
      removed = "".join(expression)
      return removed
    else:
      return None
```

The function removes punctuations from the text. It iterates over each character in the text and only keeps the character if it is not a punctuation mark.

Listing 3.5: Replace Contractions

```
@udf(StringType())
def replace_contractions(text):
    if text is not None:
        replaced = contractions.fix(text)
        return replaced.lower()
    else:
      return None
```

This function replaces contractions in the text with their expanded form. For instance, it replaces ‘it’s‘ with ‘it is‘.

Listing 3.6: Tokenize Words

```
@udf(ArrayType(StringType()))
def tokenize_words(text):
    if text is not None:
      tokenized = word_tokenize(text)
      return tokenized
    else:
      return None
```

The function tokenizes the text, which means it splits the text into individual words.

Listing 3.7: Remove Stopwords

```
@udf(ArrayType(StringType()))
def remove_stopwords(text):
    stopwords = nltk.corpus.stopwords.words('english')
    if text is not None:
      removed = [i.lower() for i in text if i not in stopwords]
      return removed
    else:
      return None
```

This function removes stopwords from the text. Stopwords are common words like ‘is‘, ‘the‘, ‘and‘, etc., that do not carry much meaning and are often removed in NLP tasks.

Listing 3.8: Lemmatize Words

```
@udf(ArrayType(StringType()))
def lemmatize_words(text):
    lemmatizer = WordNetLemmatizer()
    if text is not None:
      lemmatized = [lemmatizer.lemmatize(i) for i in text]
      return lemmatized
    else:
      return None
```

This function lemmatizes the words in the text. Lemmatization is the process of reducing a word to its base or root form. For example, ‘running‘ becomes ‘run‘, ‘better‘ becomes ‘good‘, etc.

3.3.2 Comparison of Raw and Preprocessed Data

Table 3.5 below presents a comparative view of the raw and preprocessed text data. This comparison clearly shows that the preprocessing steps have significantly changed the original text, making it more suitable for NLP tasks. The following transformations have been applied to the raw text:

- **HTML Entities Replacement:** All HTML entities in the text have been replaced with their corresponding characters. This step ensures that the text is human-readable and that no important information is lost in translation.

- **URL Removal:** Any URLs present in the text have been removed. This is done to ensure that the text data only contains relevant information for sentiment analysis.

- **Punctuation Removal:** All punctuation marks have been removed from the text. This step simplifies the text and helps in focusing on the words rather than the punctuation.

- **Stopword Removal:** Commonly used words (stopwords) that do not carry significant meaning have been removed from the text. This step helps in focusing on the important words in the text.

- **Lowercasing:** The entire text has been converted to lowercase. This ensures that the same word in different cases is not considered as different words by the NLP model.

- **Tokenization:** The text has been broken down into individual words or tokens. This is a crucial step in NLP as it allows the model to understand and analyze each word in the text separately.

Raw Text	**Preprocessed Text**
I see sustainable 20k by Q4 2021. Solid 12 months out from now. Then a nice green spring 2022.	i,see,sustainable,20k,by,q4,2021.,solid,12, month,out,from,now,then,a,nice, green,spring,2022,.
I think when btc Hit the old ath from 2019 we make another 20% fomo pump to around 15 -16k and retest the 139k as support. Btc will leave this Chanel when we hit the 2019 ath. When the bears win and defeat the Bulls After double top we see the 12k and tray a retest in the incoming weeks.	i,think,when,btc,hit,the,old,ath,from,2019,we, make,another,20,fomo,pump,to,around,15,16k, and,retest,the,139k,a,support,btc,will,leave,this, chanel,when,we,hit,the,2019,ath,when,the,bear, win,and,defeat,the,bull,after,double,top,we,see, the,12k,and,tray,a,retest, in,the,incoming,week
I have no frickin idea what's going to happen! But I will be hodling until the end!	i,have,no,frickin,idea,what,is,going,to,happen, but,i,will,be,hodling,until,the,end
If 20k is the new 10k than it's good for ETH. ETH to the moon	if,20k,is,the,new,10k,than,it,is,good,for,eth, eth,to,the,moon
Protest Elon Musk's market manipulation: http://www.youtube.com/results?search_query =%23stopelon #StopElon crypto"	protest,elon,musk,s,market,manipulation, crypto
The cycles will extend so far out that the price will seem to not move. A 10k run-up will be a bull market one day. It's not hard to see https://www.youtube.com/ watch?v=HLipKXba2Xo&t=33m40s 33:40."	the,cycle,will,extend,so,far,out,that,the,price, will,seem,to,not,move,a,10k,run,up,will,be,a, bull,market,one,day,it,is,not,hard,to,see
Benâ€™s hair signalling a STRONG V shape recovery.	ben,s,hair,signalling,a,strong,v,shape,recovery
&quot;High Risk&quot;	high,risk
I think itâ€™s going down to $12000â€¦sorry guys. Itâ€™s going to get grim.	i,think,it,is,going,down,to,12000,sorry,guy,it,is, going,to,get,grim
It's time to buy.. I will be explaining this time to my friends aftet 1-2months	it,is,time,to,buy,i,will,be,explaining,this,time, to,my,friend,aftet,1,2months

Table 3.5: Comparison of Raw and Preprocessed Text

Chapter 4

Data Selection

Within the data collected, a substantial amount of comments are found that are not related to Bitcoin or the overall cryptocurrency market. While these comments might be interesting from a broader perspective, they do not add value to the core objective of conducting sentiment analysis in the cryptocurrency market context. Hence, it is critical to discard these irrelevant comments to refine the dataset and improve the accuracy of the following analysis.

To accomplish this, a binary classifier needs to be employed that can categorize each comment into one of two categories: related or unrelated to Bitcoin and the cryptocurrency market. Although building such a classifier from scratch is feasible, the approach taken here involves using an existing, advanced model and tailoring it to this particular task. This not only conserves considerable time and resources but is also likely to yield superior results.

Considering these factors, the chosen model for this task is RoBERTa, known for being a robustly optimized BERT pretraining approach. Acknowledged for its high performance across diverse natural language processing tasks, RoBERTa is an ideal fit for this binary classification task. The subsequent sections provide a detailed account of how RoBERTa was fine-tuned for this specific task, and how it was utilized to refine the dataset.

4.1 RoBERTa as a Classifier

RoBERTa [Liu et al., 2019], which stands for Robustly Optimized BERT Pretraining Approach, is a variant of BERT (Bidirectional Encoder Representations from Transformers) that is designed to provide more robust and efficient pretraining of natural language processing models.

BERT [Devlin et al., 2019], and by extension RoBERTa, is based on the transformer architecture, which uses self-attention mechanisms to capture the context of a word in a document. The transformer model can be represented by the following formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The formula depicts the attention mechanism inherent to the transformer model. At its core, this mechanism calculates similarity measures between the query matrix Q and the key matrix K. After scaling these measures and applying a softmax function, they are transformed into weights. Lastly, these weights inform a weighted sum of the value matrix V, producing an output matrix where each row corresponds to the attention output for a specific query. This process allows the model to focus adaptively on various segments of the input, capturing intricate relationships within the data.

RoBERTa builds on BERT by adjusting key hyperparameters, removing the next-sentence pretraining objective, and training with much larger mini-batches and learning rates.

To utilize RoBERTa as a classifier, a classification layer is added on top of the RoBERTa base model. The model receives a sequence of tokens as input. These tokens are embedded into vectors and then traversed through a series of transformer blocks. The output from the final transformer block feeds into the classification layer, which then yields a probability distribution over the classes. The class securing the highest probability is selected as the prediction.

4.2 Training The RoBERTa Model

The subsequent part of this study focuses on the adaptation and utilization of the RoBERTa model for binary classification. The model's primary function is to discern and remove comments that lack relevance to Bitcoin. To optimize this task, the already pre-trained RoBERTa-Base model is refined further, tailoring it to the specific requirements of this problem.

The first step in this process involves manually assigning labels to 2000 comments, marking them either as 0 or 1. A label of 0 signifies that a comment bears no relevance to Bitcoin, while a 1 indicates its relevance to Bitcoin. A sample of Bitcoin-related comments labelled as 1 are presented in the Table 4.1 below. Examples include 'realistic,15k,18k', 'that,is,funny,because,bit,boy,said,eth,to,20k,by,end,of, the,year,lol', and 'i,sold,waiting,for,a,dip'. Conversely, comments that don't connect to the cryptocurrency market or Bitcoin carry a label of 0, examples of which are 'sound,like,an,episode,from,the,big, bang,theory', 'nice,haircut', and 'dude,you,are,awesome,love,the,channel'.

Preprocessed Text	**Label**
my,plan,is,to,follow,professor,crypto,aka,dr,benjamin	0
fully,agree,a,correction,is,inevitable,risk,v,reward,is,crazy, for,anyone,new,entering,the,space,wait,for,a,better, buying,opportunity	1
new,investor,so,cute,i,see,bitcoin,tank,and,i,jump,for,joy	1
realistic,15k,18k	1
sound,like,an,episode,from,the,big,bang,theory	0
that,is,funny,because,bit,boy,said,eth,to,20k,by,end,of, the,year,lol	1
dude,you,are,awsome,love,the,channel	0
i,sold,waiting,for,a,dip	1
nice,haircut	0

Table 4.1: Sample of Manually Labeled Comments

Upon the successful labeling of 2000 comments, the dataset was partitioned into training and testing subsets. The training subset comprised 80% (1600) of the entire data, whereas the testing subset comprised the remaining 20% (400).

To attain the optimal test accuracy, a series of experiments were conducted utilizing grid search. In total, six models were trained, each characterized by the following parameters:

- Model 1 - 15 epochs, batch size 12
- Model 2 - 15 epochs, batch size 16
- Model 3 - 20 epochs, batch size 12
- Model 4 - 20 epochs, batch size 16
- Model 5 - 25 epochs, batch size 12
- Model 6 - 25 epochs, batch size 16

The accuracies of each model are shown in the Figure 4.1 below.

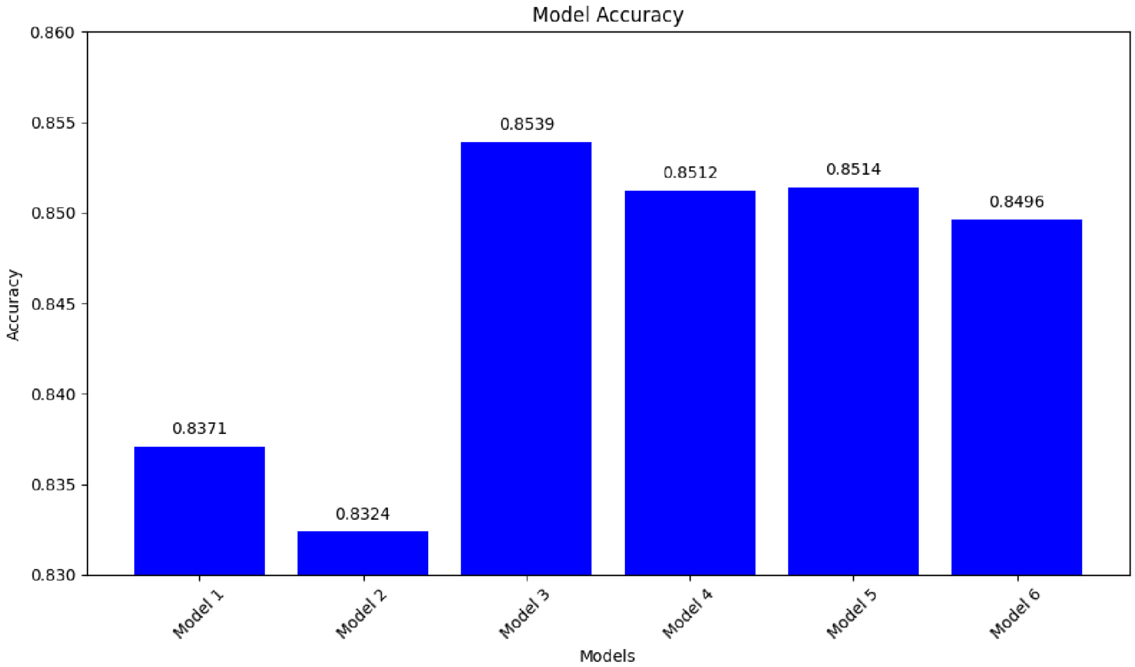


Figure 4.1: Model Accuracy

The chart reveals that Model 3 stands out with the highest test accuracy of 0.8539. For a more comprehensive understanding of Model 3's performance, a confusion matrix(Fig 4.2) was developed. In summary, the model made accurate predictions for 344 instances (183 TN + 161 TP) and erroneous predictions for 60 instances (39 FP + 21 FN). The detailed breakdown of the confusion matrix and its results are depicted below:

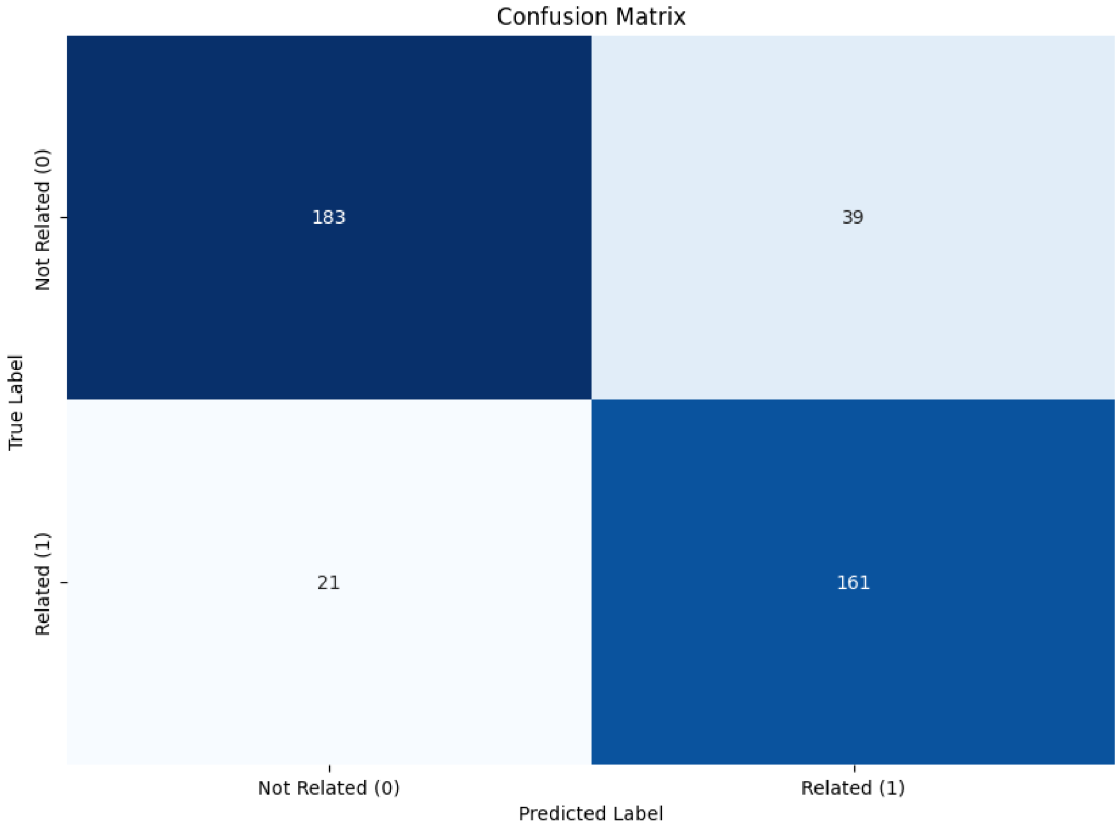


Figure 4.2: Confusion Matrix (Model 3)

True Negatives (TN): There are 183 instances where both the prediction and the actual value are "class 0". This means the model correctly predicted 183 times that a comment is not related to Bitcoin or the crypto market.

False Positives (FP): There are 39 instances where the model predicted "class 1" (related to Bitcoin or the crypto market), but the actual value was "class 0" (not related). This is a type I error.

False Negatives (FN): There are 21 instances where the model predicted "class 0" (not related), but the actual value was "class 1" (related). This is a type II error.

True Positives (TP): There are 161 instances where both the prediction and the actual value are "class 1". This means the model correctly predicted 161 times that a comment is related to Bitcoin or the crypto market.

4.3 Large-Scale Label Prediction Using The Trained RoBERTa Model

The initial dataset was composed of a total of 2,839,528 comments. After the application of the binary classification model to filter out comments not related to Bitcoin and the broader cryptocurrency market, the dataset was reduced to 1,038,265 comments that specifically pertained to these topics. It is important to note that the model used had a testing accuracy of 0.8539, indicating the potential presence of mislabeled comments. However, the ability to automate this task with an accuracy of 0.8539 is of significant value, especially when considering the alternative of manually labeling around 3 million data points.

Upon application of this model, it was discovered that a substantial portion of the comments made over the course of four years 63.4% were not focused on expressing ideas about Bitcoin or other cryptocurrencies. Instead, these comments were more focused on other aspects such as praising the content creator or commenting on the quality of the video. This insight underscores the importance and effectiveness of the binary classification model in filtering out irrelevant comments and focusing on those that provide meaningful insights into the Bitcoin and cryptocurrency market.

Chapter 5

Sentiment Analysis

5.1 TimeLM as a Sentiment Analyser

Neural language models have revolutionized the field of NLP, with models like BERT and RoBERTa leading the charge. However, these models, while powerful, have a limitation - they are essentially static, unable to adapt to the rapidly changing landscape of language use, particularly in dynamic environments like social media. To address this, a novel approach has been proposed in the form of TimeLM [Loureiro et al., 2022], a time-specific language model that is specialized for Twitter data.

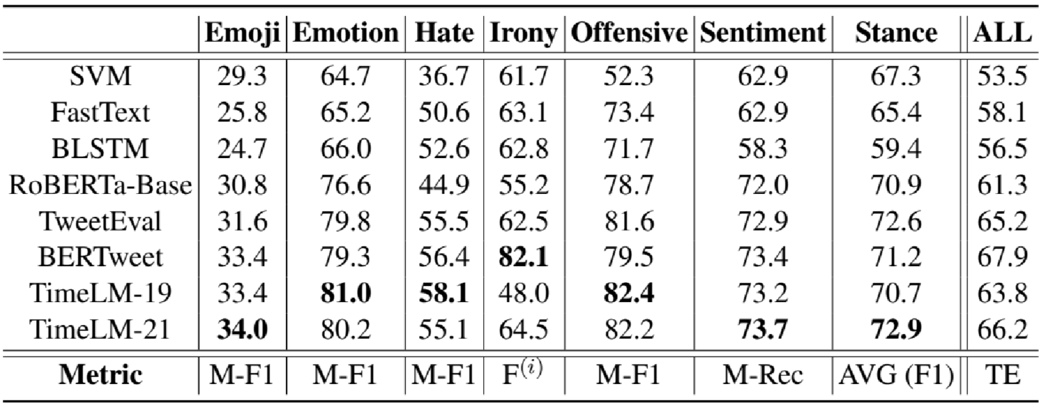
TimeLM is trained in two phases. The first phase involves training a base model using data until the end of 2019. This base model is trained using the original RoBERTa base model and continues training the masked language model on Twitter data. The choice of RoBERTa as a base model is due to its superior performance in various NLP tasks, owing to its robust architecture and the use of a dynamic masking strategy during pre-training. However, unlike RoBERTa, TimeLM is not static. It evolves over time, adapting to the changing language use on Twitter.

The second phase of TimeLM's training involves continual training of language models every three months since the date of the base model. This continual training allows TimeLM to stay up-to-date with the latest trends and topics on Twitter, making it a more accurate and relevant model for tasks involving recent data. As of 2022, the TimeLM model used for sentiment analysis has been trained on a total of 124 million tweets.

TimeLM can be used for a variety of tasks, as demonstrated by its evaluation using TweetEval [Barbieri et al., 2020], a unified Twitter benchmark composed of seven heterogeneous tweet classification tasks. These tasks include emoji prediction, emotion recognition, hate speech detection, irony detection, offensive language identification, sentiment analysis, and stance detection. For the scope of this study, the sentiment analysis of TimeLM is utilized.

The Figure 5.1 demonstrates that TimeLM-21 surpasses all other models in the task of sentiment analysis. Furthermore, when considering the average performance across all tasks, BERTweet [Nguyen et al., 2020] accomplishes a superior performance over the other models. Since the task is limited to sentiment analysis, the TimeLM-21 will be used to capitalize on its superior performance compared to other models.

It is noteworthy that these models are primarily trained on Tweets due to its convenience as a source of social sentiment. Still, these pre-trained models can effectively analyze sentiment in the YouTube comments used in this research.



	Emoji	**Emotion**	**Hate**	**Irony**	**Offensive**	**Sentiment**	**Stance**	**ALL**
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
RoBERTa-Base	30.8	76.6	44.9	55.2	78.7	72.0	70.9	61.3
TweetEval	31.6	79.8	55.5	62.5	81.6	72.9	72.6	65.2
BERTweet	33.4	79.3	56.4	**82.1**	79.5	73.4	71.2	67.9
TimeLM-19	33.4	**81.0**	**58.1**	48.0	**82.4**	73.2	70.7	63.8
TimeLM-21	**34.0**	80.2	55.1	64.5	82.2	**73.7**	**72.9**	66.2
Metric	M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG (F1)	TE

Figure 5.1: Comparison of Neural Language Models

5.2 Further Data Preprocessing

After preprocessing the data and filtering out comments related to Bitcoin and cryptocurrency using the fine-tuned RoBERTa sentiment classifier (as detailed in Chapter 4), the data is now ready to be fed into the TimeLM-21 model. However, there is an additional step that can be taken to enhance the precision of the sentiment analysis task.

In the context of social media, different keywords can often represent the same concept. For instance, the keywords ”btc” and ”bitcoin” both refer to Bitcoin, while ”eth”, ”ether”, ”etherium”, and ”ethereum” all refer to Ethereum. To improve the model's accuracy, similar keywords can be consolidated into a single representative keyword.

To identify which keywords might have a high impact, the most frequently occurring words in the dataset of 1 million comments can be counted. The Table 5.1 below shows the result of this word count:

Word	**Count**
to	547914
is	541104
i	469644
a	410372
and	408655
...	...
bitcoin	151808
btc	144797
crypto	104855
buy	104442
eth	88617
xrp	57848
ada	42888
cardano	40471

Table 5.1: Word Counts in Dataset

Upon thorough analysis, several keywords were identified for unification. The following code was used to replace these keywords in our dataset:

Listing 5.1: Regex Transformations

```
df['TEXT'] = df['TEXT'].str.replace(r'\bbtc\b', 'bitcoin', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\bbitcoins\b', 'bitcoin', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\beth\b', 'ethereum', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\betherium\b', 'ethereum', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\bether\b', 'ethereum', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\bada\b', 'cardano', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\bmusk\b', 'elon', regex=True)
df['TEXT'] = df['TEXT'].str.replace(r'\bhodl\b', 'hold', regex=True)
```

The code snippet provided above serves to consolidate various keywords into a single, unified term. For example, "btc" is transformed into "bitcoin", while "eth", "etherium", and "ether" are all standardized to "ethereum". The term "musk" is replaced with "elon", acknowledging the reference to Elon Musk, a prominent figure in the cryptocurrency discourse. Additionally, the crypto-specific term "hodl" is replaced with "hold". This term, originating from a misspelling in a forum post, has come to symbolize the strategy of maintaining ownership of Bitcoin or Ethereum, despite market fluctuations. Merging these terms strengthens the data's consistency, leading to increased accuracy in the following analysis.

5.3 Word Cloud - A Visual Representation of Mentions

A word cloud is used to display a list of words, the importance of each being shown with font size or color. This format is useful for quickly perceiving the most prominent terms in a large corpus of text and it is a popular way to depict keyword metadata that accompanies text data.

In the context of this study, now that the data has been thoroughly cleaned and transformed, a word cloud can be generated to gain a better understanding of the main topics being discussed in the dataset of 1 million comments. This visual tool allows for quick identification of the most frequently used words, providing an overview of the prevailing themes in the discourse.

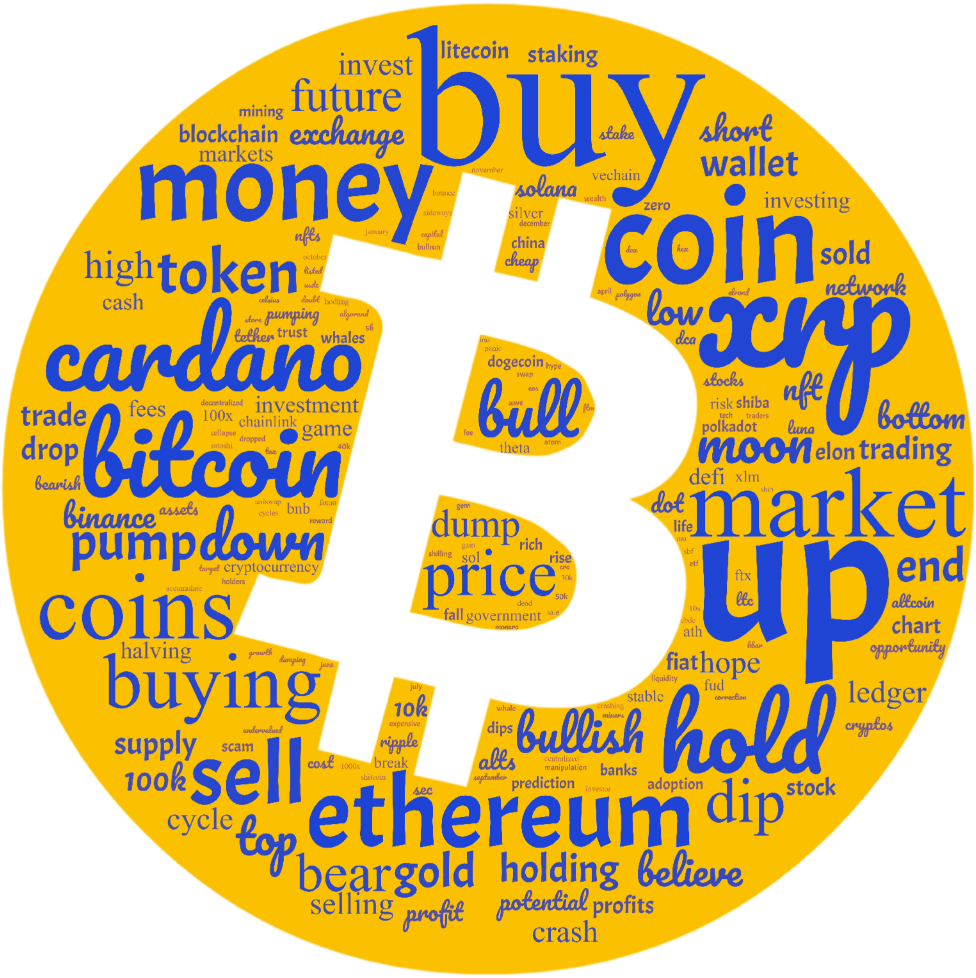


Figure 5.2: Word Cloud

5.4 Implementation of TimeLM

The TimeLM-21 model, specifically fine-tuned for sentiment analysis, along with other TimeLM-21 models, are readily available on the Hugging Face model platform. To utilize these models, the Transformers library [Face, 2023] in Python is employed, which provides easy access to the Hugging Face models. The model used for sentiment analysis is identified as follows:

```
#Define model version
model_name = "cardiffnlp/twitter-roberta-base-sentiment-latest"
#Instantiate the pipeline object
classifier = pipeline(’sentiment-analysis’, model=model_name)
```

Given the extensive size of the dataset, encompassing 1,038,265 comments, and the computational demands of sentiment score calculation, using a distributed framework becomes crucial. PySpark is employed for this task, and a UDF is created to calculate the sentiment score for each comment.

```
@udf(FloatType())
def get_sentiment_score(comment):
    result = classifier(comment)[0]
    if result[’label’] == "positive":
      return result[’score’]
    elif result[’label’] == "negative":
      return result[’score’]*(-1)
    else:
      return result[’score’]+1
```

Inherently, TimeLM outputs a sentiment score and label as distinct outcomes. Yet, calculating these results independently can be resource-intensive. To streamline this, the UDF is adjusted to combine these steps. When the label is positive, the function directly returns the score. For a negative label, the score is multiplied by (-1). In cases of a neutral label, 1 is added to the score. Utilizing this method, scores are directly acquired and subsequently categorized as required.

To finish, this UDF is applied to the ’TEXT’ column to derive the sentiment scores with the subsequent command:

```
dfspark = dfspark.withColumn("SCORE", get_sentiment_score("TEXT"))
```

This step enriches the dataset with sentiment scores, paving the way for further analysis.

ID	TEXT	LIKES	SCORE	LABEL
232334	cardano will be big	11	0.8844477	positive
232369	the bottom is 0 remember i called it	2	0.7043731	neutral
125893	nope 2023 will be worst	3	0.89353925	negative
38175	10k bears are the new 100k bulls	0	0.6907114	neutral
126080	just holding coins is also not profitable	0	0.79998416	negative
3245	the epic dumpage is days away 6k is thr final bottom so sorry	2	0.87437075	negative
38672	if binance goes we all go the best fud is the best buy signal	47	0.72853243	positive
128980	i am super bullish	1	0.96792555	positive
10655	is it just me or do these youtubers claim rare signal flash every other day	48	0.8318274	neutral

Table 5.2: Sample of Dataset with Sentiment Scores

The Table 5.2 above provides a snapshot of the dataset after sentiment scores have been calculated. Each row represents a comment, with the 'TEXT' column containing the comment text, the 'LIKES' column indicating the number of likes the comment received, the 'SCORE' column showing the sentiment score calculated by our model, and the 'LABEL' column displaying the sentiment label (positive, neutral, or negative).

From the provided sample, a diverse range of sentiments becomes evident in the comments, with sentiment scores corresponding to the anticipated sentiment of the comment text. For instance, the comment "cardano will be big" yields a positive sentiment score of 0.8844477, tagged as 'positive'. On the other hand, the remark "nope 2023 will be worst" registers a negative sentiment score of 0.89353925, tagged as 'negative'. The coherence between sentiment scores and labels indicates the model's effective performance in designating sentiment to the comments.

Sentiment	Count
negative	248788
neutral	542038
positive	247439

Table 5.3: Count of Sentiment Scores For Each Label

From the Table 5.3, it is clear that the dataset is heavily skewed towards 'neutral' sentiments, accounting for 542,038 instances. This suggests that a large part of the data under review conveys a neutral sentiment.

'Negative' and 'positive' sentiments appear almost equally in the data, with counts of 248,788 and 247,439, respectively. This implies a balance in the data between these two sentiments, and a strong presence of neutral sentiment.

Chapter 6

Preliminary Data Exploration

After obtaining sentiment scores for the entire dataset using the TimeLM sentiment analysis model, the next step is to delve deeper into the data.

The initial analysis phase groups the data on a daily basis. This approach determines the count of negative, positive, and neutral comments for each day spanning from January 1, 2019, to January 1, 2023. Such grouping helps identify potential fluctuations or trends in sentiment over time and provides insight into the daily distribution of sentiments.

For a more comprehensive analysis, not only are daily sentiments counted, but Bitcoin prices for each corresponding date are also included. The Table 6.1 below showcases the combined dataset for the first ten days.

Date	Negative	Neutral	Positive	Total	BTC Price
2019-01-01	35	63	30	128	3746.71
2019-01-02	19	41	26	86	3849.22
2019-01-03	17	56	55	128	3931.05
2019-01-04	42	110	93	245	3832.04
2019-01-05	35	77	45	157	3851.97
2019-01-06	12	40	32	84	3836.52
2019-01-07	41	128	66	235	4078.58
2019-01-08	24	81	38	143	4028.47
2019-01-09	37	75	52	164	4031.55
2019-01-10	26	61	34	121	4034.41

Table 6.1: Sample of Daily Sentiment Distribution

By analyzing this combined dataset, I aim to uncover any potential relationships or correlations between public sentiment and Bitcoin's market performance, and thereby shed light on the influence of public mood on cryptocurrency markets.

6.1 Descriptive Analysis

In the subsequent descriptive analysis, key statistical parameters such as the minimum and maximum values, quartiles, and standard deviations for each sentiment category are examined.

An interesting observation is the close proximity of the average values for positive and negative sentiments, which are 169.36 and 170.29, respectively. However, the average number of neutral comments noticeably exceeds these values, implying a more prevalent neutral sentiment in the collected data.

At its peak, the count of negative comments reached a striking high of 1617, significantly overshadowing the maximum count of positive comments, which only reached 787. This suggests that there was a particular day during the four-year period under study where the sentiment was decidedly negative.

Table 6.2: Negative

Metric	Value
Mean	170.29
Std. Deviation	143.91
Min	6
25% (Q1)	71
Median (Q2)	130
75% (Q3)	235
Max	1617

Table 6.3: Neutral

Metric	Value
Mean	371.00
Std. Deviation	296.37
Min	19
25% (Q1)	177
Median (Q2)	275
75% (Q3)	492
Max	2106

Table 6.4: Positive

Metric	Value
Mean	169.36
Std. Deviation	124.62
Min	6
25% (Q1)	86
Median (Q2)	133
75% (Q3)	225
Max	787

6.2 Distribution of The Sentiment Groups

Constructing a histogram for each sentiment category and supplementing it with a Kernel Density Estimation (KDE) offers valuable insights into the distribution of comment counts within each group. This visual approach aids in discerning patterns, trends, and anomalies with clarity. The Figure 6.2 illustrates these distributions.

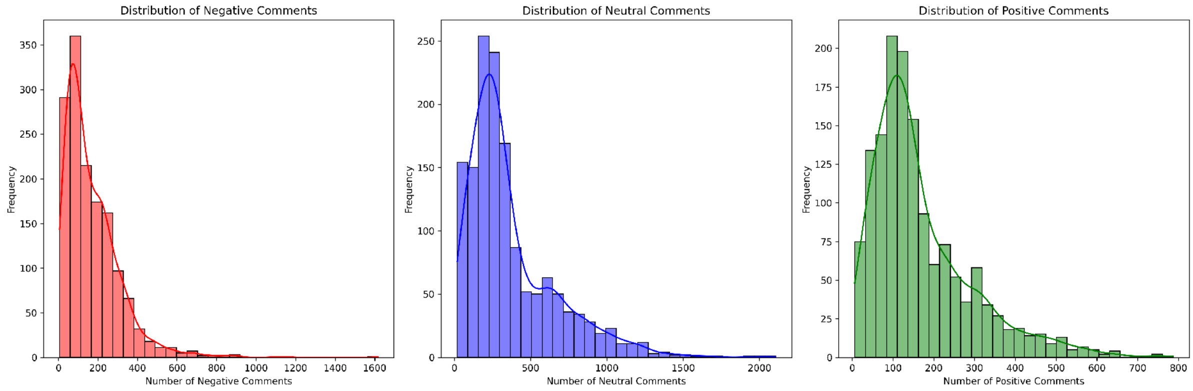


Figure 6.1: Distribution of Each Sentiment Group

The distribution of negative comments is positively skewed with the majority of data points centered around lower values. This indicates that on most days, the number of negative comments was relatively low. However, there were also some days with a significantly higher number of negative comments.

Similar to negative comments, the distribution of neutral comments is positively skewed. The majority of the data points are situated around the lower values, indicating that the number of neutral comments was relatively low on most days. There were, however, some days with a significantly higher number of neutral comments.

The distribution of positive comments is also positively skewed, with most of the data points gathered around the lower values. This indicates that on most days, the number of positive comments was relatively low. Nevertheless, there were a few days with a significantly higher number of positive comments.

While all three categories show similar patterns, it is clear that there is a significant difference in the distribution of counts between negative, neutral, and positive comments. The neutral comments, for instance, appear to have a broader spread and higher maximum count compared to the negative and positive comments.

Along with the histogram, another useful graphical representation would be the cumulative distribution of the sentiment categories. This type of plot can provide additional insights into the distribution and progression of the sentiment scores.

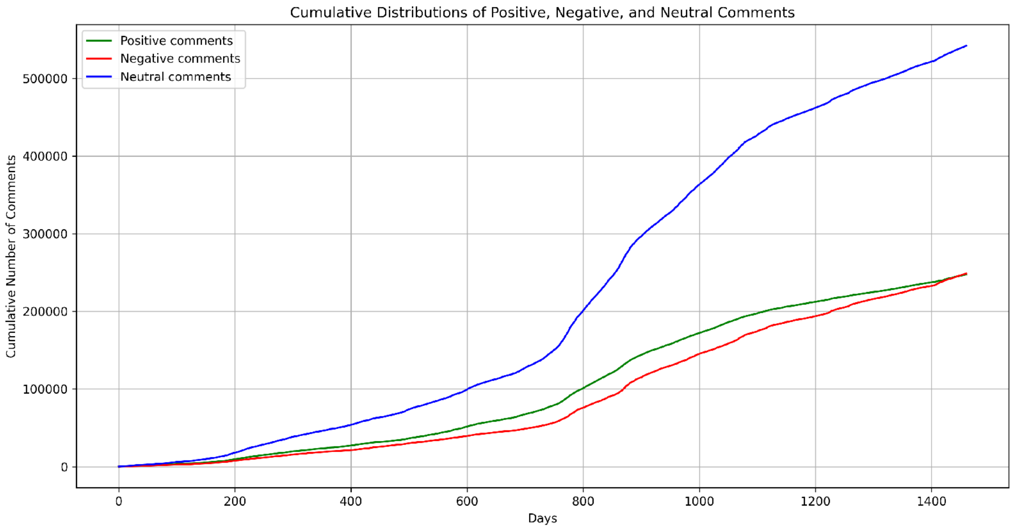


Figure 6.2: Cumulative Distribution of Each Sentiment Group

The plot above shows the cumulative distributions of positive (green), negative (red), and neutral (blue) comments over time.

The x-axis represents the date, while the y-axis represents the cumulative number of comments. Each point on a line represents the total number of comments of that type up to and including that date.

It can be observed that the cumulative number of each type of comment increases over time, as expected. Interestingly, the cumulative number of neutral comments seems to be consistently higher than that of positive or negative comments, suggesting that neutral comments might be more common overall.

Chapter 7

Discovering The Relationship

7.1 Total Comment Count vs. Bitcoin Price Over Time

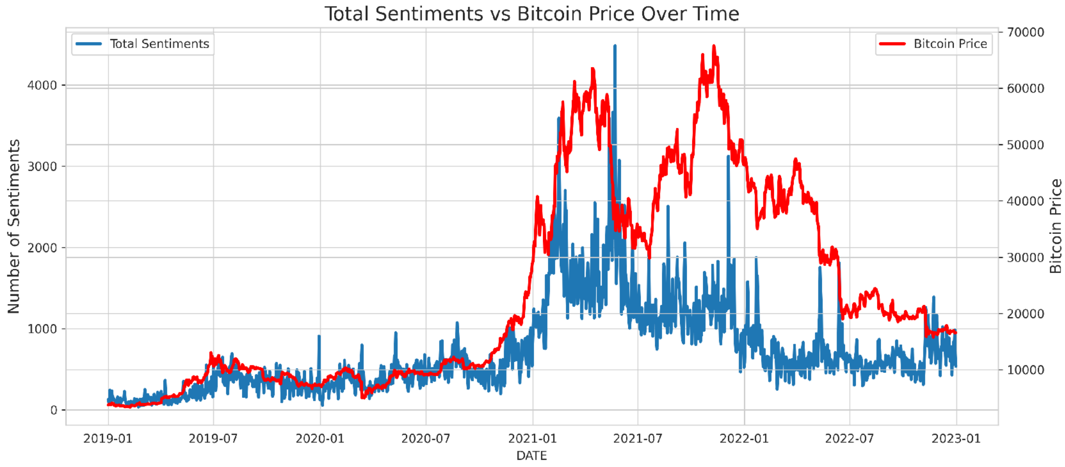


Figure 7.1: Overlay of Total Sentiment Count and Bitcoin Price Over Time

The visual (Figure 7.1) reveals a concurrent trend between the overall sentiment count and the Bitcoin price, suggesting a strong relationship between these two variables. Specifically, periods of heightened engagement within the YouTube communities—marked by an increased volume of comments—appear to coincide with increases in Bitcoin's price. Conversely, when the level of engagement diminishes, indicated by a decrease in the number of comments, we observe a corresponding drop in Bitcoin's price.

There are notable instances where a significant drop in Bitcoin's price has spurred a temporary surge in engagement, as reflected in the comment count. These spikes likely represent periods of intense discussion triggered by the dramatic price movements.

The Pearson correlation coefficient between the total sentiment count and the Bitcoin price is approximately 0.75. This strong positive correlation implies that as the total number of sentiments (comprising positive, negative, and neutral comments) rises, the Bitcoin price tends to increase correspondingly. The extremely small p-value (almost zero) reaffirms the statistical significance of this correlation, suggesting that the observed relationship is unlikely to be due to random chance.

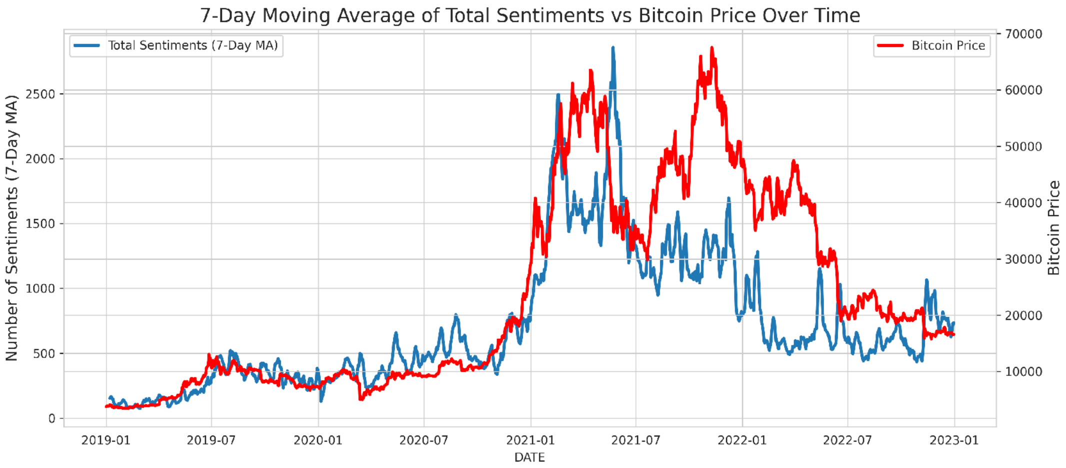


Figure 7.2: Overlay of Total Sentiment Count(7-Day MA) and Bitcoin Price Over Time

To provide a clearer view of the underlying trends and minimize the effects of daily variations, a 7-day moving average was applied to the total sentiment count (Figure 7.2). This method enhances the understanding of the connection between public sentiment and Bitcoin price dynamics across the timeframe.

7.2 Net Sentiment (Positive-Negative) vs. Bitcoin Price Over Time

To examine the impact of the difference between positive and negative sentiment on BTC price, the following graph (Figure 7.3) can be referenced.

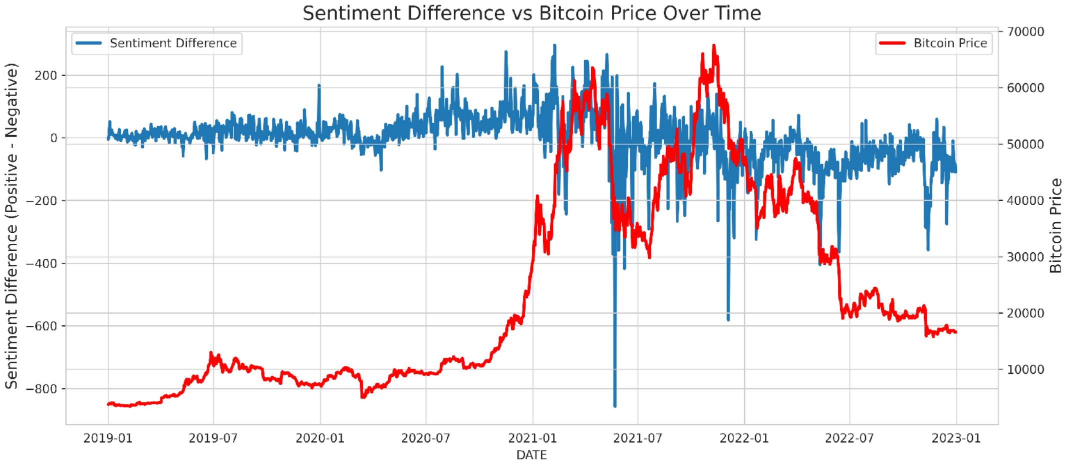
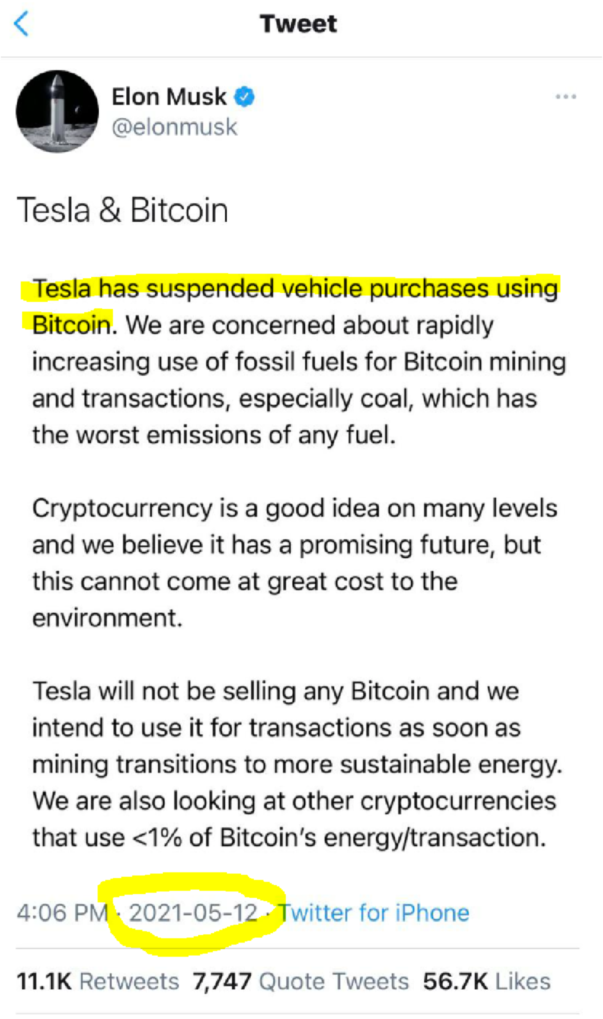


Figure 7.3: Overlay of Net Sentiment Count and Bitcoin Price Over Time

The blue line in the graph represents the difference between positive and negative sentiments each day. Positive values indicate days when positive comments outnumbered negative ones, and vice versa. The red line represents the Bitcoin price each day.

Upon examining the plot, it is clear that periods when the sentiment difference dips below zero often coincide with substantial declines in Bitcoin's price. This pattern is particularly pronounced when the sentiment difference exceeds -800 in May 2021, corresponding with a dramatic drop in Bitcoin's value, often referred to as the ”Elon Musk crash.” This event followed a tweet by Elon Musk, stating that Tesla would no longer accept Bitcoin for car purchases due to environmental concerns, which led to widespread market turmoil. This downturn occurred shortly after Bitcoin reached an all-time high of approximately $65,000 in April 2021.



Elon Musk
@elonmusk

Tesla & Bitcoin

Tesla has suspended vehicle purchases using Bitcoin. We are concerned about rapidly increasing use of fossil fuels for Bitcoin mining and transactions, especially coal, which has the worst emissions of any fuel.

Cryptocurrency is a good idea on many levels and we believe it has a promising future, but this cannot come at great cost to the environment.

Tesla will not be selling any Bitcoin and we intend to use it for transactions as soon as mining transitions to more sustainable energy. We are also looking at other cryptocurrencies that use <1% of Bitcoin's energy/transaction.

4:06 PM · 2021-05-12 · Twitter for iPhone

11.1K Retweets 7,747 Quote Tweets 56.7K Likes

Figure 7.4: Elon Musk's Tweet

Another notable dip in sentiment difference, around -500, is evident in November 2021. This decline in sentiment occurred immediately after Bitcoin achieved its second all-time high of $69,000 (the current record as of this analysis). The plot shows that Bitcoin's price continued to fall in the days following this sentiment drop, despite brief upticks during relief rallies.

In essence, the chart reveals a pattern: each sharp downward movement in Bitcoin's price corresponds to a similarly sharp decline in sentiment difference. This trend suggests a potential correlation between sentiment difference and Bitcoin price, particularly during periods of market volatility.

While the graphical representation of the data suggests a discernible pattern, it is important to validate these observations and investigate any possible causal relationships. To achieve this, a Lead-Lag analysis will be conducted using the concept of Granger Causality.

7.3 Lead-Lag Analysis

Lead-Lag Analysis is a technique used to determine the sequencing of price movements between two or more variables. A leading variable is one that makes its move before the other variables do, while a lagging variable is one that reacts to the movement of the leading variable. In the context of this study, a lead-lag relationship would imply that changes in sentiment difference precede changes in Bitcoin price (or vice versa).

On the other side, Granger Causality [Granger, 1969] is a statistical hypothesis test used to determine whether one time series is useful in forecasting another. In essence, a time series X is said to Granger-cause Y if it can be shown, usually through a series of t-tests and F-tests, that the probabilities of Y occurring are altered by the past values of X. It is worth noting that Granger causality is not 'causality' in the traditional sense; rather, it is a useful statistical tool to identify predictive relationships.

To run the Granger Causality test on the dataset, it is crucial to confirm that the time series data in question is stationary. This is a fundamental prerequisite because the Granger Causality test operates on the premise that the time series under examination remain stationary.

Stationarity refers to a property of a time series whereby its statistical properties, such as mean and variance, remain constant over time. Essentially, a stationary time series does not exhibit trends or seasonality and its fluctuations around the mean are consistent over time.

To determine the stationarity of the time series data, the Augmented Dickey-Fuller (ADF) test can be employed. This test offers a robust statistical method to evaluate stationarity. If the series are found to be non-stationary, they may need to be transformed before applying the Granger Causality test.

7.3.1 Augmented Dickey Fuller Test

The null and alternative hypotheses of the ADF test [Dickey and Fuller, 1979] are defined as follows:

Null Hypothesis (H0): The time series has a unit root, i.e., it is non-stationary. It exhibits a systematic pattern (such as a trend or seasonality) that causes the statistical properties (like the mean and variance) to change over time.

Alternative Hypothesis (H1): The time series does not have a unit root, i.e., it is stationary. The statistical properties of the series are constant over time, meaning the series does not exhibit trends or seasonality that cause these properties to change.

Upon utilizing the Augmented Dickey-Fuller (ADF) test on the sentiment difference and Bitcoin price series, the resulting p-values are as follows:

Sentiment Difference: The obtained p-value, 0.0054, falls beneath the standard 0.05 threshold for statistical significance. Therefor the null hypothesis can be rejected, which posits that the sentiment difference series is non-stationary. Consequently, the sentiment difference series appears to be stationary, without evident trends or seasonal patterns influencing its statistical attributes over its duration.

Bitcoin Price: The obtained p-value is 0.5358, exceeding the standard 0.05 threshold. Thus, the null hypothesis isn't rejected, indicating that the Bitcoin price series might be non-stationary. As expected this suggests that there are inherent patterns in the Bitcoin price series, like trends or seasonal variations, which impact its statistical behavior over time. Before proceeding with the Granger Causality test, this series will require transformation to achieve stationarity.

7.3.2 Ensuring Stationarity

To ensure that the Bitcoin price series is stationary, a technique known as 'differencing' is applied to the time series.

Differencing is a method used to eliminate trends and seasonality. It operates by subtracting the preceding observation from the current observation. In mathematical notation, differencing can be expressed as:

$$\text{diff}(t) = x(t) - x(t-1)$$

where $x(t)$ represents the value of the time series at time t, and diff(t) denotes the differenced series.

After differencing the Bitcoin price series and reapplying the Augmented Dickey-Fuller (ADF) test, a p-value of approximately 4.98×10^{-11} is obtained, which is significantly below 0.05. This result rejects the null hypothesis of the ADF test, suggesting that the differenced series is stationary. Thus, the differenced Bitcoin price series is confirmed to be stationary.

With the sentiment difference series and the differenced Bitcoin price series both verified as stationary, it becomes feasible to proceed with the Granger Causality Test. This analysis will shed light on potential lead-lag relationships between these two series.

7.3.3 Granger Causality

Granger causality is a statistical concept of causality that is based on prediction. According to this concept, if a signal X_t "Granger-causes" (or "G-causes") a signal Y_t, then past values of X_t should contain information that helps predict Y_t above and beyond the information contained in past values of Y_t alone.

The Granger causality test is a hypothesis test to determine if one time series is useful in forecasting another. While it is called 'causality', it is more accurately described as a test of whether the future values of one variable are predicted by the past values of another, which is not the same as causality in the usual sense.

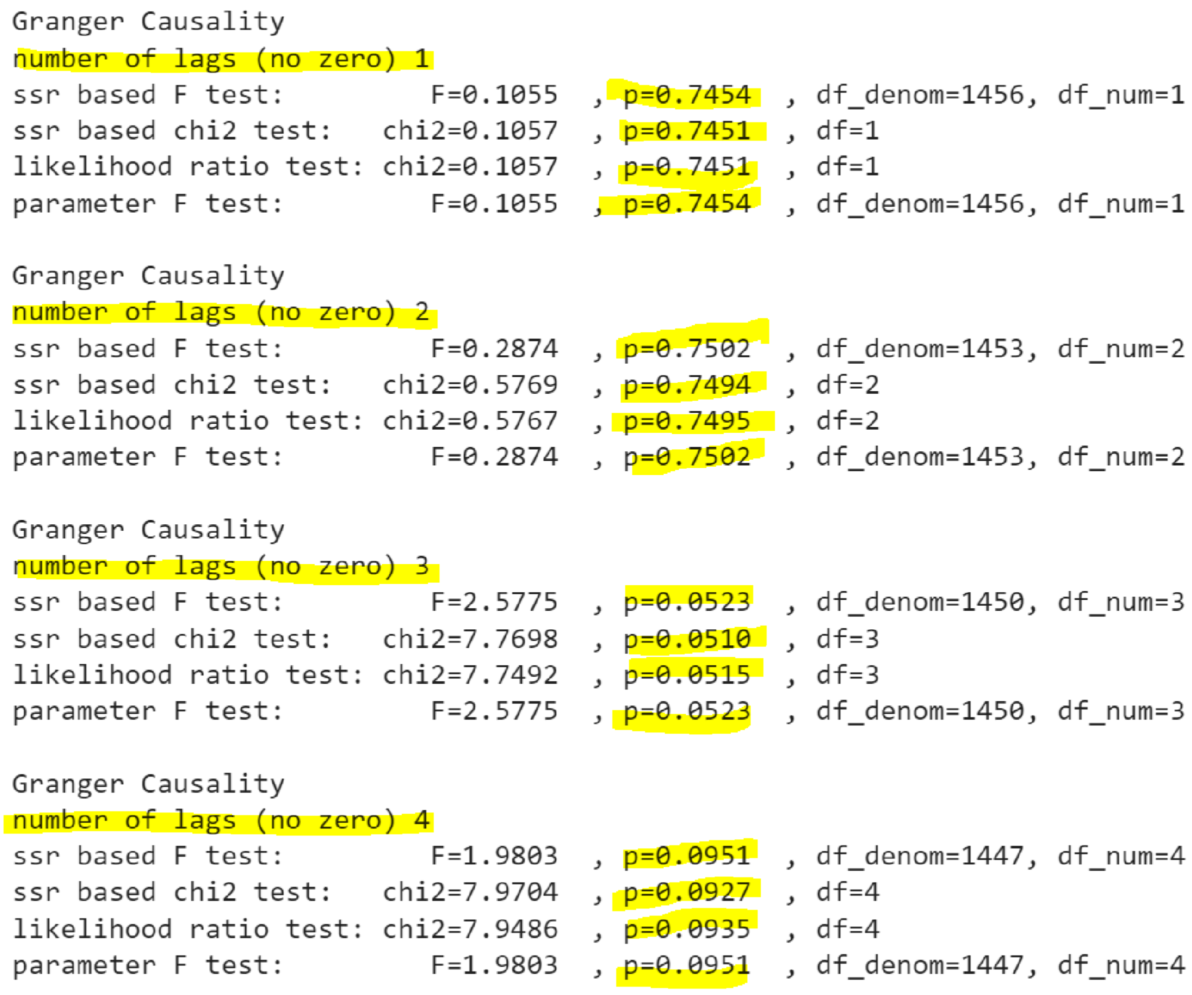
The null and alternative hypotheses of the Granger causality test are as follows:

Null Hypothesis (H0): The coefficients of the past values in the regression equation are zero, i.e., X_t does not Granger-cause Y_t. In simpler terms, this means that the past values of X_t have no effect on the current value of Y_t.

Alternative Hypothesis (H1): The coefficients of the past values are not zero, i.e., X_t does Granger-cause Y_t. This means that the past values of X_t do affect the current value of Y_t.

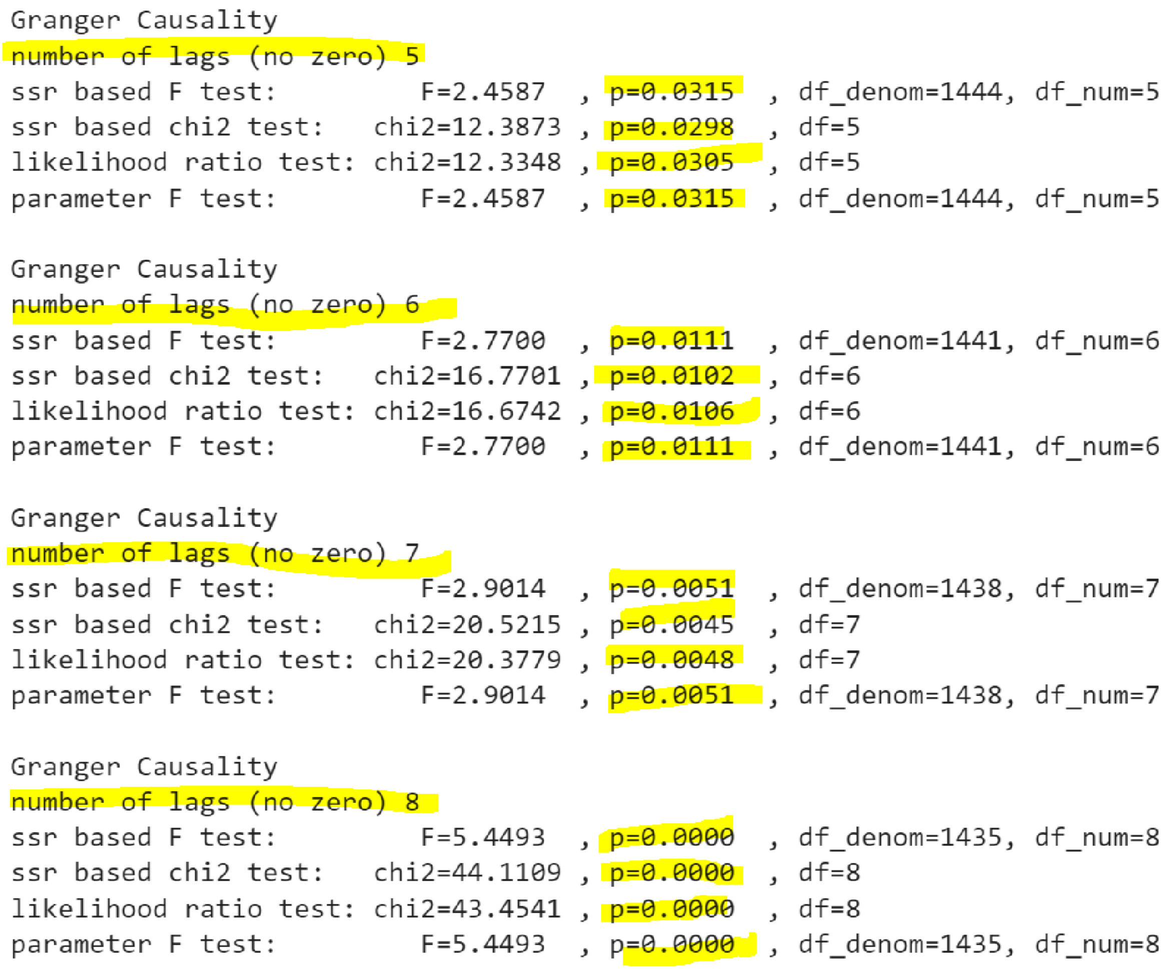
If the p-value obtained from the Granger causality test is below a the determined significance level ,0.05, then the null hypothesis can be rejected and conclude that the past values of X_t do have predictive power on Y_t, i.e., X_t does Granger-cause Y_t.

Below are the Granger-Causality test results as follows :



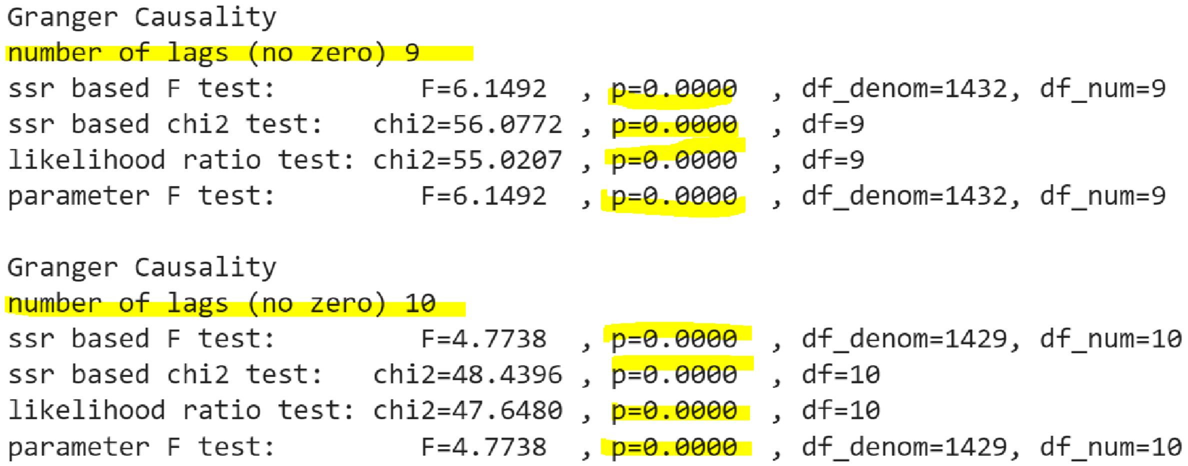
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.1055  , p=0.7454  , df_denom=1456, df_num=1
ssr based chi2 test:   chi2=0.1057  , p=0.7451  , df=1
likelihood ratio test: chi2=0.1057  , p=0.7451  , df=1
parameter F test:         F=0.1055  , p=0.7454  , df_denom=1456, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.2874  , p=0.7502  , df_denom=1453, df_num=2
ssr based chi2 test:   chi2=0.5769  , p=0.7494  , df=2
likelihood ratio test: chi2=0.5767  , p=0.7495  , df=2
parameter F test:         F=0.2874  , p=0.7502  , df_denom=1453, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=2.5775  , p=0.0523  , df_denom=1450, df_num=3
ssr based chi2 test:   chi2=7.7698  , p=0.0510  , df=3
likelihood ratio test: chi2=7.7492  , p=0.0515  , df=3
parameter F test:         F=2.5775  , p=0.0523  , df_denom=1450, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=1.9803  , p=0.0951  , df_denom=1447, df_num=4
ssr based chi2 test:   chi2=7.9704  , p=0.0927  , df=4
likelihood ratio test: chi2=7.9486  , p=0.0935  , df=4
parameter F test:         F=1.9803  , p=0.0951  , df_denom=1447, df_num=4
```

Figure 7.5: Granger Causality Test (Part1)



```
Granger Causality
number of lags (no zero) 5
ssr based F test:         F=2.4587  , p=0.0315  , df_denom=1444, df_num=5
ssr based chi2 test:   chi2=12.3873 , p=0.0298  , df=5
likelihood ratio test: chi2=12.3348 , p=0.0305  , df=5
parameter F test:         F=2.4587  , p=0.0315  , df_denom=1444, df_num=5

Granger Causality
number of lags (no zero) 6
ssr based F test:         F=2.7700  , p=0.0111  , df_denom=1441, df_num=6
ssr based chi2 test:   chi2=16.7701 , p=0.0102  , df=6
likelihood ratio test: chi2=16.6742 , p=0.0106  , df=6
parameter F test:         F=2.7700  , p=0.0111  , df_denom=1441, df_num=6

Granger Causality
number of lags (no zero) 7
ssr based F test:         F=2.9014  , p=0.0051  , df_denom=1438, df_num=7
ssr based chi2 test:   chi2=20.5215 , p=0.0045  , df=7
likelihood ratio test: chi2=20.3779 , p=0.0048  , df=7
parameter F test:         F=2.9014  , p=0.0051  , df_denom=1438, df_num=7

Granger Causality
number of lags (no zero) 8
ssr based F test:         F=5.4493  , p=0.0000  , df_denom=1435, df_num=8
ssr based chi2 test:   chi2=44.1109 , p=0.0000  , df=8
likelihood ratio test: chi2=43.4541 , p=0.0000  , df=8
parameter F test:         F=5.4493  , p=0.0000  , df_denom=1435, df_num=8
```

Figure 7.6: Granger Causality Test (Part2)



```
Granger Causality
number of lags (no zero) 9
ssr based F test:         F=6.1492  , p=0.0000  , df_denom=1432, df_num=9
ssr based chi2 test:   chi2=56.0772 , p=0.0000  , df=9
likelihood ratio test: chi2=55.0207 , p=0.0000  , df=9
parameter F test:         F=6.1492  , p=0.0000  , df_denom=1432, df_num=9

Granger Causality
number of lags (no zero) 10
ssr based F test:         F=4.7738  , p=0.0000  , df_denom=1429, df_num=10
ssr based chi2 test:   chi2=48.4396 , p=0.0000  , df=10
likelihood ratio test: chi2=47.6480 , p=0.0000  , df=10
parameter F test:         F=4.7738  , p=0.0000  , df_denom=1429, df_num=10
```

Figure 7.7: Granger Causality Test (Part3)

The Granger Causality Test was performed for up to 10 lags. The results indicate that for the initial four lags, the p-values exceed 0.05, suggesting that sentiment differences do not Granger-cause variations in Bitcoin price. However, beginning from the 5th lag and continuing onward, the p-values decrease and fall below 0.05 from the 8th lag, suggesting a rejection of the null hypothesis. This implies that historical sentiment differences appear to have some predictive power over changes in Bitcoin price. The predictive power becomes substantially stronger from the 8th lag onwards, where the p-value is virtually 0.

To use 8 lags for predicting Bitcoin price, features consisting of sentiment differences from the past 8 days should be generated. This means, for predicting the Bitcoin price on a specific day, sentiment differences from that day and the 7 preceding days are taken as features.

It is crucial to understand that ”Granger causality” does not equate to actual causality. The observation that sentiment differences Granger-cause Bitcoin price fluctuations does not confirm that sentiment shifts directly influence Bitcoin price changes. Instead, it suggests that sentiment differences can serve as a valuable predictor for Bitcoin price shifts. The actual causal relationship might be more complicated and may involve other factors that are not included in this particular analysis.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

In this study, I explored the potential of an often-underestimated resource, YouTube, as a supplementary metric for predicting Bitcoin price fluctuations. Specifically, I sourced comments from selected influential YouTube channels' cryptocurrency-related videos to conduct a sentiment analysis. This analysis aimed to gauge sentiment shifts over a four-year span, from 2019 to 2023. Subsequently, I examined the correlation between Bitcoin's price trends and the evolving sentiment during this period to ascertain the predictive value of YouTube comments on Bitcoin's future price movements. A Granger causality test was applied to these two time series, and the findings were evaluated based on the F-statistic and its corresponding p-value.

In conclusion, the Granger causality analysis reveals that sentiments derived from YouTube comments can serve as a valuable addition to existing predictive metrics in cryptocurrency literature. This test was conducted with a lag length of 10, with the F-statistic peaking at the 9th lag length. Significantly, the associated p-value was below the customary 0.05 threshold, suggesting that sentiment scores, based on the past 9 days, Granger-cause Bitcoin's price variations. In simpler terms, a nine-day historical sentiment score can aid in forecasting Bitcoin's price trajectory.

However, it is crucial to highlight that Bitcoin's price dynamics are multifaceted and cannot be accurately predicted using a singular metric. Numerous factors, including supply and demand dynamics, regulatory interventions, technological advancements, market manipulations by major stakeholders (often termed "whales"), halving events, liquidity considerations, and more, influence Bitcoin's price. This research endeavor has illuminated the potential of social media sentiment, especially from YouTube, as a valuable predictor. Yet, it is just one facet of the broader forecasting framework. Integrating this sentiment metric into a comprehensive predictive model will offer a more rounded perspective. By doing so, we can better determine the true impact and enhancement this research brings to the overarching predictive accuracy of Bitcoin's price trajectory.

8.2 Future Work

There are three distinct dimensions in this study that present opportunities for enhancement and further integration. These are detailed as follows:

Improving the Binary Classification Model: The binary classification model employed in this study, while effective, has room for enhancement. The model was trained on a relatively small dataset of 2,000 manually labeled data points to categorize the remaining 3 million comments. With an accuracy of 0.8539, there is potential for refining its precision. By incorporating a more extensive manually labeled training dataset, the model's accuracy can be bolstered, leading to more nuanced results in the sentiment analysis phase.

Integration into a Comprehensive Predictive Model: The parameters and insights unearthed in this study can be seamlessly integrated as features into a more intricate predictive model tailored for Bitcoin price prediction. By doing so, the true efficacy of the sentiment metric, as a predictor, can be rigorously tested and validated.

Real-time Sentiment Analysis Pipeline: Once the model's positive impact on Bitcoin price prediction is ascertained, the next logical progression would be the implementation of a data pipeline. This pipeline would capture real-time shifts in sentiment, potentially on a daily basis. By harnessing real-time data, the model can be continually updated, ensuring it remains relevant and accurate in the ever-evolving landscape of cryptocurrency.

By addressing these avenues in future research endeavors, the bridge between social media sentiment and cryptocurrency price prediction can be further strengthened, offering valuable insights to stakeholders in the cryptocurrency domain.

Bibliography

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Cryptography Mailing list at https://metzdowd.com*, 03 2009.

Anamika, Madhumita Chakraborty, and Sowmya Subramaniam. Does sentiment impact cryptocurrency? *Journal of Behavioral Finance*, 24 (2):202–218, 2023. doi: 10.1080/15427560.2021.1950723. URL `https://doi.org/10.1080/15427560.2021.1950723`.

Zi Ye, Yinxu Wu, Hui Chen, Yi Pan, and Qingshan Jiang. A stacking ensemble deep learning model for bitcoin price prediction using twitter comments on bitcoin. *Mathematics*, 10(8), 2022. ISSN 2227-7390. doi: 10.3390/math10081307. URL `https://www.mdpi.com/2227-7390/10/8/1307`.

Huali Zhao., Martin Crane., and Marija Bezbradica. Attention! transformer with sentiment on cryptocurrencies price prediction. In *Proceedings of the 7th International Conference on Complexity, Future Information Systems and Risk - COMPLEXIS*, pages 98–104. INSTICC, SciTePress, 2022. ISBN 978-989-758-565-4. doi: 10.5220/0011103400003197.

Erdinc Akyildirim, Ahmet Faruk Aysan, Oguzhan Cepni, and S. Pinar Ceyhan Darendeli. Do investor sentiments drive cryptocurrency prices? *Economics Letters*, 206:109980, 2021. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2021.109980. URL `https://www.sciencedirect.com/science/article/pii/S0165176521002573`.

Jacques Vella Critien, Albert Gatt, and Joshua Ellul. Bitcoin price change and trend prediction through twitter sentiment and data volume. *Journal of Financial Innovation*, 8, 05 2022. doi: 10.1186/s40854-022-00352-7.

David Bourghelle, Fredj Jawadi, and Philippe Rozin. Do collective emotions drive bitcoin volatility? a triple regime-switching vector approach.

Journal of Economic Behavior and Organization, 196:294–306, 04 2022. doi: 10.1016/j.jebo.2022.01.026.

Mehmet Balcilar, Elie Bouri, Rangan Gupta, and David Roubaud. Can volume predict bitcoin returns and volatility? a quantiles-based approach. *Economic Modelling*, 64:74–81, 2017. ISSN 0264-9993. doi: https://doi.org/10.1016/j.econmod.2017.03.019. URL `https://www.sciencedirect.com/science/article/pii/S0264999317304558`.

Niranjan Sapkota. News-based sentiment and bitcoin volatility. *International Review of Financial Analysis*, 82(C), 2022. doi: 10.1016/j.irfa.2022.10218. URL `https://ideas.repec.org/a/eee/finana/v82y2022ics1057521922001454.html`.

Khaled Mokni, Ahmed Bouteska, and Mohamed Sahbi Nakhli. Investor sentiment and bitcoin relationship: A quantile-based analysis. *The North American Journal of Economics and Finance*, 60:101657, 2022. ISSN 1062-9408. doi: https://doi.org/10.1016/j.najef.2022.101657. URL `https://www.sciencedirect.com/science/article/pii/S1062940822000171`.

Xiaoquan Jiang, Iván M. Rodríguez, and Qianying Zhang. Macroeconomic fundamentals and cryptocurrency prices: A common trend approach. *Financial Management*, 52(1):181–198, March 2023. doi: 10.1111/fima.12412. URL `https://ideas.repec.org/a/bla/finmgt/v52y2023i1p181-198.html`.

Arianna Agosto, Paola Cerchiello, and Paolo Pagnottoni. Sentiment, google queries and explosivity in the cryptocurrency market. *Physica A: Statistical Mechanics and its Applications*, 605:128016, 2022. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2022.128016. URL `https://www.sciencedirect.com/science/article/pii/S0378437122006380`.

Sandy Suardi, Atiqur Rahman Rasel, and Bin Liu. On the predictive power of tweet sentiments and attention on bitcoin. *International Review of Economics and Finance*, 79:289–301, 2022. ISSN 1059-0560. doi: https://doi.org/10.1016/j.iref.2022.02.017. URL `https://www.sciencedirect.com/science/article/pii/S1059056022000375`.

Sotirios Oikonomopoulos, Katerina Tzafilkou, Dimitrios Karapiperis, and Vassilios Verykios. Cryptocurrency price prediction using social media sentiment analysis, 2022.

Junwei Chen. Analysis of bitcoin price prediction using machine learning. *Journal of Risk and Financial Management*, 16(1), 2023. ISSN 1911-8074. doi: 10.3390/jrfm16010051. URL `https://www.mdpi.com/1911-8074/16/1/51`.

YouTube. Youtube data api, 2023. URL `https://developers.google.com/youtube/v3`.

ranaroussi. Yahoo finance api, 2023. URL `https://github.com/ranaroussi/yfinance`.

The Apache Software Foundation. Pyspark, 2023. URL `https://github.com/apache/spark/tree/master/python/pyspark`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. Timelms: Diachronic language models from twitter, 2022.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets, 2020.

Hugging Face. Hugging face, transformers python library, 2023. URL `https://github.com/huggingface/transformers`.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1912791`.

D. Dickey and Wayne Fuller. Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74, 06 1979. doi: 10.2307/2286348.