Analyzing Sentiment Trends and Patterns in Bitcoin-Related Tweets Using TF-IDF Vectorization and K-Means Clustering

Tri Wahyuningsih[1,*], Shih Chih Chen[2]

[1]Doctorate Program of Computer Science, Universitas Kristen Satya Wacana, Jawa Tengah, Indonesia

[2]Department of Information Management, National Kaohsiung University of Science and Technology, Taiwan

ABSTRACT

This study conducts a comprehensive analysis of Bitcoin-related tweets to understand sentiment trends and patterns using TF-IDF vectorization and K-means clustering. The dataset, comprising 1,544 unique tweets, was collected via the Twitter API and preprocessed to remove duplicates and clean the text. Sentiment analysis revealed a distribution of 53.7% neutral, 29.7% positive, and 16.6% negative tweets, indicating a predominant neutral sentiment in the discourse. Keyword analysis identified frequent terms such as 'bitcoin' (479 occurrences), 'new' (46), 'good' (43), 'crypto' (39), and 'trade' (39). Visualizations through word clouds highlighted the specific language associated with each sentiment category, with positive tweets focusing on opportunities and innovation, while negative tweets emphasized risks and scams. Cluster analysis using K-means, with the optimal number of clusters determined by the elbow method, resulted in three distinct clusters. Cluster 0, comprising 1,346 tweets, was characterized by neutral and informative content, focusing on market updates and trading strategies. Cluster 1, with 163 tweets, contained a higher concentration of positive sentiment, highlighting positive developments and investment opportunities. Cluster 2, the smallest with 35 tweets, focused on negative sentiment, reflecting concerns about market volatility and fraudulent activities. These clusters provided a nuanced understanding of the thematic composition of Bitcoin-related tweets. The study's findings have practical implications for investors, traders, and market analysts by providing insights into market mood and sentiment trends. The integration of these findings into predictive models can enhance market prediction accuracy and develop more effective trading strategies. Despite the study's contributions, limitations such as the dataset's language and scope suggest areas for future research, including real-time sentiment analysis and the incorporation of multimodal data sources. This research advances the field of sentiment analysis in financial markets, particularly within the context of cryptocurrencies, by offering a detailed and longitudinal examination of social media sentiment.

Keywords Bitcoin, sentiment analysis, TF-IDF vectorization, K-means clustering, social media, Twitter, cryptocurrency, market sentiment, keyword analysis, cluster analysis

Submitted 10 January 2024
Accepted 20 April 2024
Published 1 June 2024

Corresponding author
Tri Wahyuningsih,
982022001@student.uksw.edu

Additional Information and Declarations can be found on page 67

DOI: 10.47738/jcrb.v1i1.11

INTRODUCTION

In recent years, the influence of social media on financial markets has become increasingly apparent. Platforms like Twitter, Facebook, and Reddit have evolved beyond their initial roles as social networking sites to become significant sources of real-time information and sentiment analysis. Investors, analysts, and traders frequently monitor social media channels to gauge public sentiment

How to cite this article: T. Wahyuningsih and S. C. Chen, "Analyzing Sentiment Trends and Patterns in Bitcoin-Related Tweets Using TF-IDF Vectorization and K-Means Clustering," *J. Curr. Res. Blockchain*, vol. 1, no. 1, pp. 48-69, 2024.

and make informed decisions. This shift has led to the development of sophisticated tools and methodologies aimed at extracting actionable insights from social media data. Sentiment analysis, a subfield of natural language processing (NLP), has emerged as a crucial technique for understanding public opinion and its potential impact on market dynamics.

Bitcoin, the first and most well-known cryptocurrency, exemplifies the volatile nature of digital assets. Since its inception in 2009 by an anonymous entity known as Satoshi Nakamoto, Bitcoin has experienced significant fluctuations in value, driven by a myriad of factors including regulatory news, technological advancements, market adoption, and public sentiment. The decentralized nature of Bitcoin, combined with its potential for high returns, has attracted a diverse group of investors ranging from individual enthusiasts to large institutional players. This volatility, while offering substantial profit opportunities, also entails considerable risk, making the ability to predict price movements highly valuable. Understanding how social media sentiment affects Bitcoin's price is thus not only academically intriguing but also practically significant for market participants.

Social media, particularly platforms like Twitter, plays a significant role in shaping public opinion. Researchers have explored how social media serves as a representation of public opinion [1]. Moreover, research has delved into how social media platforms like Twitter can shape public opinion during events like elections [2].

Analyzing discussions on social media, especially Twitter, provides a unique opportunity to understand public sentiment and opinions on various topics [3]. social media has been proven effective in activating and mobilizing public opinion, as seen during election campaigns [4]. While Twitter users may not be fully representative of the general population, their expressions on the platform still contribute to the overall formation of public opinion [5].

Twitter can be a valuable tool for predicting stock price movements and understanding investor behavior. Studies have shown that sentiment analysis of Twitter data can predict movements in stock indices [6]. The daily happiness index, constructed through sentiment analysis of Twitter messages, is used to study the impact of social media on financial markets [7]. Sentiment analysis has been recognized as a significant factor in predicting financial markets due to the reflection of market trends in social media activity.

With its real-time information dissemination capabilities, Twitter has become a vital tool for investors and traders who seek to understand the collective mood and opinions about various market entities, including cryptocurrencies. The platform allows for quick sharing and spreading of news, opinions, and rumors, which can significantly influence the perceptions and behaviors of market participants. High-profile figures, such as CEOs, financial analysts, and influential social media personalities, often use Twitter to express their views, which can cause immediate and substantial reactions in the market. The decentralized and open nature of Twitter means that information, whether accurate or speculative, can rapidly affect public sentiment and, consequently, market dynamics.

Previous research has consistently highlighted the significant impact of social media sentiment on Bitcoin prices. Studies have shown that positive tweets

about Bitcoin can lead to a surge in its price, while negative sentiment can cause sharp declines. For instance, [8] demonstrated that Twitter mood states could be correlated with the stock market movements, suggesting a similar application could be extended to Bitcoin. Similarly, [9] found that social media activity could be a predictor of Bitcoin's price volatility. These studies underline the importance of monitoring social media sentiment as a part of a comprehensive market analysis strategy. However, the rapidly evolving nature of social media and the increasing volume of data necessitate ongoing research to refine sentiment analysis techniques and improve predictive accuracy. This paper aims to build upon existing literature by providing a detailed analysis of Bitcoin-related tweets, identifying trends and patterns in sentiment, and exploring their potential implications for Bitcoin market dynamics.

Despite the growing body of research on the impact of social media sentiment on financial markets, there remains a notable lack of comprehensive studies specifically analyzing the sentiment of Bitcoin-related tweets over extended periods. Most existing studies tend to focus on short-term sentiment analysis or isolated events, leaving a significant gap in understanding how sentiment trends evolve over time. This gap is crucial because the sentiment expressed in tweets can fluctuate widely with news events, regulatory announcements, and broader market trends. Without a longitudinal approach, it is challenging to discern whether observed sentiment changes are transient or indicative of longer-term trends.

Understanding trends and patterns in Bitcoin-related sentiment is essential for several reasons. First, it can provide insights into the collective mood of market participants, which is a key driver of market behavior. Identifying patterns can help differentiate between short-lived market reactions and more sustained shifts in sentiment. Second, a comprehensive analysis can reveal how sentiment correlates with market movements, potentially enhancing predictive models for Bitcoin prices. Finally, by uncovering the underlying drivers of sentiment, such as specific types of news or influential tweets, stakeholders can better anticipate and respond to market changes. This study aims to address this research gap by providing a detailed, longitudinal analysis of Bitcoin-related tweets, focusing on identifying significant trends and patterns and exploring their implications for the cryptocurrency market.

The primary objective of this research is to analyze the sentiment distribution of Bitcoin-related tweets. By examining the proportion of positive, neutral, and negative sentiments expressed in tweets, this study aims to provide a clear picture of the general mood surrounding Bitcoin over time. This analysis will not only offer insights into the prevailing sentiment but also help in understanding how sentiment correlates with major market events and news. A thorough sentiment distribution analysis can reveal patterns and trends that may be indicative of broader market sentiments, contributing to a deeper understanding of the market's emotional dynamics.

Another key objective is to identify common keywords and phrases associated with each sentiment category. By performing a detailed keyword analysis, this study will uncover the specific language and terms that are frequently used in positive, neutral, and negative tweets. This analysis will involve extracting and examining keywords and phrases to determine their association with different sentiments. Identifying these keywords will help in understanding the drivers of

sentiment and provide insights into the topics and themes that influence public perception of Bitcoin. This understanding is crucial for market analysts and investors who rely on sentiment analysis to gauge market mood and predict potential price movements.

Lastly, the study aims to group tweets into clusters based on sentiment and content to uncover common themes and patterns. By using clustering techniques, tweets will be categorized into distinct groups that share similar characteristics. This approach will help in identifying major themes and narratives within the Bitcoin discourse on Twitter. Clustering will not only enhance the understanding of sentiment but also reveal the underlying topics that dominate discussions about Bitcoin. These insights can be valuable for identifying emerging trends, understanding public concerns, and predicting market reactions to different types of information. By achieving these objectives, this research will contribute to the growing body of knowledge on the role of social media sentiment in financial markets, particularly in the context of cryptocurrencies.

Literature Review

Sentiment Analysis in Financial Markets

Sentiment analysis, a subfield of natural language processing (NLP), involves the computational study of opinions, sentiments, and emotions expressed in text. In financial markets, sentiment analysis has become a crucial tool for understanding investor behavior and market dynamics. By analyzing large volumes of textual data from news articles, social media posts, and financial reports, sentiment analysis can provide insights into the collective mood of market participants. These insights can, in turn, influence trading strategies, risk management, and investment decisions. The basic premise is that the emotions and opinions expressed in public forums can serve as indicators of future market movements.

The applications of sentiment analysis in financial markets are diverse and growing. One of the primary uses is in predicting stock prices. Researchers have developed models that correlate the sentiment extracted from social media posts and news articles with stock market performance. Positive sentiment generally correlates with upward price movements, while negative sentiment often precedes price declines. Additionally, sentiment analysis is used to gauge market reactions to earnings announcements, regulatory changes, and geopolitical events. By providing a real-time understanding of market sentiment, these models help investors anticipate market shifts and make more informed decisions.

Several key studies have explored the impact of social media sentiment on financial markets, particularly stock prices and cryptocurrencies. Research by [8] demonstrated the predictive power of Twitter mood states on the dow jones industrial average (DJIA). Their research showed that certain mood indicators derived from Twitter data could predict stock market movements with a high degree of accuracy. This study laid the groundwork for subsequent research into the relationship between social media sentiment and financial markets.

In the context of cryptocurrencies, [9] explored the relationship between Bitcoin prices and online search queries, finding that increased search activity often

preceded price increases. Similarly, [10] examined the role of social signals in Bitcoin markets, demonstrating that social media activity could explain significant variations in Bitcoin trading volumes and price volatility. More recently, studies like those by [11] have highlighted how social media sentiment can serve as an early warning system for market movements, providing valuable signals for traders and investors. These studies collectively underscore the importance of integrating sentiment analysis into financial market models, particularly in the highly volatile and sentiment-driven cryptocurrency markets.

Sentiment Analysis Techniques

Sentiment analysis techniques can be broadly categorized into two main approaches: lexicon-based methods and machine learning-based methods. Lexicon-based techniques rely on predefined lists of words (lexicons) that are associated with positive, negative, or neutral sentiments. These methods evaluate the sentiment of a text by counting the occurrences of sentiment-laden words and calculating an overall sentiment score. Examples of popular sentiment lexicons include SentiWordNet, AFINN, and the NRC Emotion Lexicon. Lexicon-based methods are relatively straightforward to implement and interpret, making them a popular choice for initial sentiment analysis tasks.

Machine learning-based techniques, on the other hand, involve training models on labeled datasets to classify text into different sentiment categories. These methods leverage algorithms such as Naive Bayes, Support Vector Machines (SVM), and more recently, deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models learn to recognize patterns in the data and can often achieve higher accuracy than lexicon-based methods, particularly when dealing with large and complex datasets. Additionally, machine learning models can be fine-tuned on domain-specific data, making them highly adaptable to different contexts and applications.

In comparing the effectiveness of lexicon-based and machine learning-based sentiment analysis methods, several studies have highlighted the strengths and limitations of each approach. Lexicon-based methods are generally faster and require less computational power, making them suitable for real-time analysis and applications where interpretability is crucial. However, these methods can struggle with context and nuance, often missing the subtleties of sarcasm, irony, or domain-specific jargon. For instance, studies have shown that while lexicon-based methods perform adequately on general text, their accuracy diminishes when applied to specialized fields like finance or medicine without proper customization.

Machine learning-based methods, conversely, excel in handling complex language patterns and large datasets. Research has demonstrated that these methods consistently outperform lexicon-based approaches in terms of accuracy and robustness. For example, a study found that machine learning models, particularly those utilizing deep learning architectures, achieved higher precision and recall in sentiment classification tasks across various domains [12]. However, these methods require substantial amounts of labeled data for training and can be computationally intensive. They also tend to function as "black boxes," providing little insight into the decision-making process, which can be a drawback in scenarios where explainability is essential.

Overall, the choice between lexicon-based and machine learning-based methods depends on the specific requirements of the task at hand. While lexicon-based methods offer simplicity and speed, machine learning-based methods provide greater accuracy and adaptability, especially for complex sentiment analysis tasks. Recent advancements in hybrid approaches that combine the strengths of both techniques are also gaining traction, offering promising results in the evolving field of sentiment analysis.

Twitter and Cryptocurrency Markets

The relationship between Twitter sentiment and cryptocurrency markets has been a focal point of numerous academic studies, reflecting the growing recognition of social media's influence on financial assets. Researchers have utilized various methodologies to analyze how sentiment extracted from tweets can predict or explain cryptocurrency price movements. One notable study examined the correlation between Twitter sentiment and Bitcoin prices, employing a sentiment analysis algorithm to categorize tweets and analyze their temporal relationship with Bitcoin's market performance [13]. The findings indicated a significant correlation, suggesting that positive sentiment often precedes price increases, while negative sentiment can signal impending declines.

Another influential study investigated the impact of Twitter sentiment on multiple cryptocurrencies, including Bitcoin, Ethereum, and Ripple [14]. Using machine learning techniques to classify tweet sentiment, the researchers found that social media sentiment could indeed predict short-term price movements, particularly during periods of high market volatility. Their analysis revealed that the sentiment of influential accounts, such as those belonging to well-known industry figures or cryptocurrency influencers, had a more pronounced effect on market behavior. This study underscored the importance of considering the source and reach of tweets in sentiment analysis.

The findings from these studies highlight the complex interplay between social media sentiment and cryptocurrency markets. One consistent observation across multiple studies is that the cryptocurrency market, especially Bitcoin, is highly susceptible to sentiment-driven price movements. For instance, research by [15] demonstrated that spikes in Twitter activity, particularly those containing positive sentiment, often lead to increased trading volumes and subsequent price surges. This phenomenon is partly attributed to the decentralized and speculative nature of cryptocurrencies, where investor behavior is heavily influenced by market sentiment and news events.

Moreover, research has shown that the sentiment expressed by influential Twitter accounts can amplify market reactions. A study highlighted how tweets from key opinion leaders could significantly impact Bitcoin's market dynamics [11]. The study found that tweets from influential accounts were more likely to generate widespread attention and drive substantial price changes, suggesting a form of social media-driven market manipulation or hype. These findings emphasize the need for investors and analysts to monitor social media channels closely, not just for sentiment but also for the influence and reach of the content being disseminated.

Research Gap and Motivation

Despite the substantial body of research exploring the influence of social media sentiment on financial markets, significant gaps remain, particularly concerning the comprehensive analysis of Bitcoin-related tweets over extended periods. Many existing studies have focused on short-term sentiment analysis, often linked to specific events or brief timeframes, thereby missing the broader trends and long-term patterns that could provide deeper insights into market behavior. Additionally, while several studies have examined the relationship between social media sentiment and cryptocurrency prices, there is a paucity of research that delves into the clustering of sentiment and the thematic content of tweets. This gap is crucial because understanding the recurring themes and the evolution of sentiment over time can reveal underlying drivers of market sentiment that short-term analyses overlook.

Another notable gap in the literature is the limited exploration of the predictive power of sentiment clusters. While numerous studies have shown a correlation between aggregate sentiment and market movements, few have investigated how different clusters of sentiment—such as those associated with specific topics or events—might uniquely influence market behavior. Furthermore, the interaction between sentiment expressed by influential accounts and the broader market sentiment remains underexplored. These gaps highlight the need for a more nuanced and comprehensive approach to sentiment analysis that incorporates clustering techniques and considers the temporal dynamics of social media sentiment.

The current study aims to address these gaps by providing a detailed, longitudinal analysis of Bitcoin-related tweets, focusing on identifying significant trends and patterns in sentiment and exploring their implications for the cryptocurrency market. By leveraging advanced sentiment analysis and clustering techniques, this research will uncover common themes and the evolution of sentiment over time, offering a more holistic view of the social media landscape surrounding Bitcoin. This approach will not only enhance the understanding of how sentiment influences Bitcoin prices but also provide insights into the underlying factors driving market sentiment.

Moreover, the study's focus on clustering tweets based on sentiment and content represents a novel contribution to the field. By identifying and analyzing clusters of tweets, this research will reveal how different themes and topics correlate with market behavior, potentially uncovering new predictors of market movements. Additionally, by considering the influence of tweets from key opinion leaders, this study will shed light on the role of influential accounts in shaping market sentiment and driving price changes. These contributions are particularly relevant given the increasing importance of social media in financial markets and the growing interest in cryptocurrencies as both investment assets and subjects of academic inquiry.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in figure 1 outlines the detailed steps of the research method.

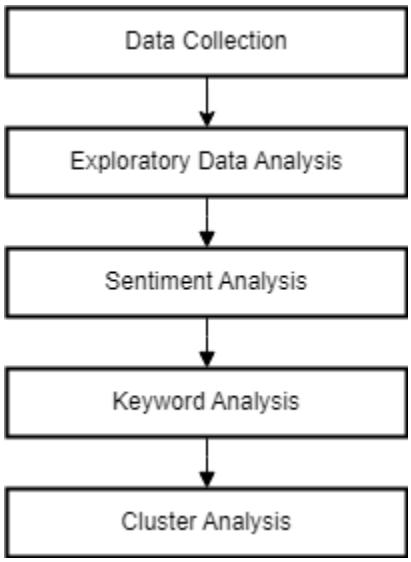

Figure 1 **Research Method Flowchart**

Data Collection

The dataset used in this study comprises Bitcoin-related tweets, each annotated with sentiment labels indicating whether the sentiment expressed is positive, neutral, or negative. This dataset provides a rich source of real-time public sentiment data, reflecting the opinions and emotions of a diverse group of Twitter users regarding Bitcoin. The data includes various features such as the tweet text, user information, and timestamps, which allow for comprehensive analysis of both the content and the temporal aspects of sentiment expression.

The tweets were collected using the Twitter API, which provides access to a vast stream of public tweets. The dataset spans a significant period, covering several months to ensure a robust analysis of sentiment trends over time. To ensure the quality and relevance of the data, specific keywords and hashtags related to Bitcoin were used as filters during the data collection process. These keywords include terms like "Bitcoin," "BTC," "cryptocurrency," and related hashtags that are commonly used in discussions about Bitcoin on Twitter. The sentiment labels were assigned using a combination of automated sentiment analysis tools and manual annotation to ensure accuracy and reliability.

The primary source of the dataset is the Twitter API, which allows for the extraction of public tweets based on specified criteria. Using the Tweepy library in Python, we streamed tweets that contained predefined Bitcoin-related keywords and hashtags. This real-time data collection approach ensured that the dataset was up-to-date and reflective of current public sentiment. The tweets were then stored in a structured format, including metadata such as the tweet ID, timestamp, user ID, and tweet text.

Before analysis, several preprocessing steps were undertaken to clean and prepare the data. Initially, duplicate tweets and retweets were removed to prevent redundancy and ensure that each tweet represented a unique sentiment expression. Following this, standard text preprocessing techniques were applied, including the removal of URLs, special characters, and stop words, as well as tokenization and lemmatization of the tweet text. These steps

were essential to normalize the text data and facilitate accurate sentiment analysis and keyword extraction. Additionally, the sentiment labels were verified and corrected where necessary through manual inspection to maintain the dataset's integrity.

Overall, these meticulous data collection and preprocessing steps were crucial to building a reliable dataset for sentiment analysis. By ensuring the data's relevance, accuracy, and cleanliness, we laid a solid foundation for subsequent analysis, enabling us to extract meaningful insights into Bitcoin-related sentiment on Twitter.

Exploratory Data Analysis (EDA)

The initial step in our exploratory data analysis involved thorough data preprocessing to ensure the dataset's quality and suitability for further analysis. First, we addressed the issue of duplicate entries. Duplicate tweets, which can arise from retweets or multiple postings of the same content, were identified and removed. This step was crucial to prevent the overrepresentation of certain sentiments or opinions, which could bias the analysis.

Next, we handled missing values, though our dataset fortunately had no missing entries, as indicated by the complete counts across all columns. Following this, we performed text cleaning, an essential step in preparing textual data for analysis. This process included tokenization, which breaks down the tweet text into individual words or tokens, making it easier to analyze the content. Additionally, we removed stop words—common words like "and," "the," and "is" that do not carry significant meaning in sentiment analysis. Special characters, URLs, and other non-alphanumeric elements were also stripped from the text to further standardize the data. These preprocessing steps were vital in transforming raw tweet text into a clean and analyzable format.

Following data preprocessing, we conducted an initial analysis to gain insights into the dataset's characteristics. We started by generating summary statistics, which provided a foundational understanding of the data. The dataset consisted of 1,897 unique tweets, each annotated with one of three sentiment labels: positive, neutral, or negative. The distribution of these sentiments was as follows: 860 tweets were labeled as neutral, 779 as positive, and 258 as negative. This distribution indicated a predominant neutral sentiment in the dataset.

We also examined the length of tweets to understand the variability in content. The analysis of tweet lengths revealed that the average tweet length was approximately 90 characters, with a standard deviation of 35. The shortest tweet in the dataset was 7 characters long, while the longest tweet contained 148 characters. The interquartile range (IQR) indicated that 50% of the tweets had lengths between 61 and 119 characters, suggesting a moderate level of variability in tweet length. This information is crucial for understanding the nature of the tweets and their potential impact on sentiment analysis.

To visualize the sentiment distribution, we created bar charts as shown in figure 2, which clearly illustrated the predominance of neutral tweets, followed by positive and negative tweets. This visual representation provided a quick and intuitive understanding of the sentiment landscape within the dataset.

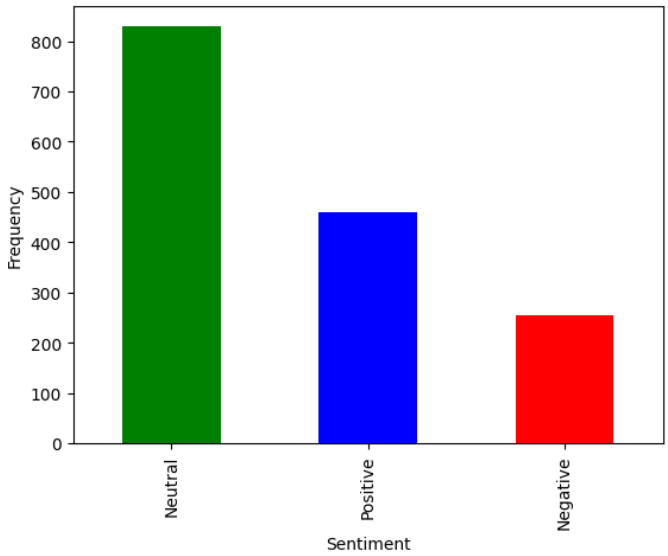

Figure 2 **Sentiment Distribution**

Additionally, we generated histograms as shown in figure 3 to depict the distribution of tweet lengths, further highlighting the variability in the data.

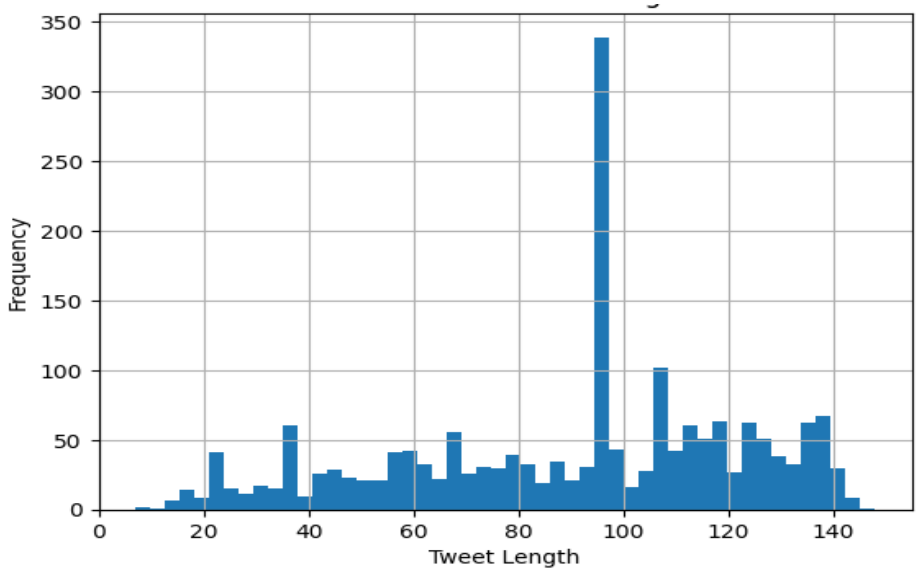

Figure 3 **Distribution of Tweet Lengths**

These initial analyses laid the groundwork for more in-depth exploratory and inferential analyses, setting the stage for uncovering trends and patterns in Bitcoin-related sentiment on Twitter.

Sentiment Analysis

The sentiment distribution analysis is a crucial step in understanding the overall mood and opinions expressed in Bitcoin-related tweets. By categorizing each tweet into positive, neutral, or negative sentiments, we can gain insights into the general public's perception of Bitcoin. In our dataset, we observed a total of 1,544 unique tweets, each labeled with one of the three sentiment categories.

Specifically, the dataset comprised 829 tweets labeled as neutral, 459 as positive, and 256 as negative. This distribution indicates that neutral sentiment is the most prevalent, followed by positive and negative sentiments.

To analyze the sentiment distribution, we calculated the proportion of each sentiment category relative to the total number of tweets. The results revealed that 53.7% of the tweets were neutral, 29.7% were positive, and 16.6% were negative. These proportions highlight a significant inclination towards neutrality in the discourse surrounding Bitcoin on Twitter. The relatively high percentage of neutral tweets suggests that while there are strong opinions expressed, many users share content that is neither explicitly positive nor negative. This neutrality could be attributed to informative posts, updates, or news articles that do not necessarily evoke a strong emotional response.

To provide a clear visual representation of the sentiment distribution, we employed bar charts, which are effective in illustrating categorical data. The bar chart for sentiment distribution visually reinforces the numerical findings, with a prominent bar representing neutral sentiment, followed by positive and negative sentiments. This visualization aids in quickly conveying the overall sentiment landscape to readers and highlights the dominance of neutral tweets in the dataset. By presenting the data in this manner, we ensure that the distribution of sentiments is easily understandable, allowing for a more intuitive grasp of the sentiment trends within the Bitcoin-related Twitter discourse.

This sentiment distribution analysis lays the groundwork for deeper investigations into how different sentiments correlate with Bitcoin market movements and other external factors. Understanding the proportion of positive, neutral, and negative sentiments is essential for developing predictive models and for further exploring the impact of sentiment on Bitcoin's price volatility. By establishing a baseline of sentiment distribution, we can better interpret subsequent analyses and draw more informed conclusions about the relationship between social media sentiment and cryptocurrency markets.

Keyword Analysis

Feature extraction is a critical step in our keyword analysis, enabling us to break down the tweet texts into manageable components for further examination. The process begins with tokenization, which involves splitting the tweet text into individual words or tokens. This step is essential for converting the raw text data into a format suitable for frequency analysis and other forms of textual analysis. Following tokenization, we perform frequency analysis to identify the most common words used in the tweets. This analysis helps us understand the key terms and topics that are frequently mentioned in relation to Bitcoin.

The frequency analysis of our dataset revealed that the most common keywords include 'bitcoin' (479 occurrences), 'new' (46 occurrences), 'good' (43 occurrences), 'crypto' (39 occurrences), 'trade' (39 occurrences), 'mining' (39 occurrences), 'btc' (38 occurrences), 'interested' (37 occurrences), 'gmail.com' (36 occurrences), and 'contact' (35 occurrences). These results indicate that discussions around Bitcoin are heavily focused on general mentions of the cryptocurrency, trading activities, and related topics such as mining and new developments. The presence of contact information like 'gmail.com' and 'contact' suggests a significant number of tweets are promotional or solicitative in nature.

To visually represent the common keywords identified in our frequency analysis, we created word clouds as shown in figure 4. Word clouds are graphical representations where the size of each word corresponds to its frequency in the text, providing an intuitive and immediate visual summary of the most prominent terms. By generating separate word clouds for positive, neutral, and negative sentiments, we can better understand the specific language and topics associated with each sentiment category.

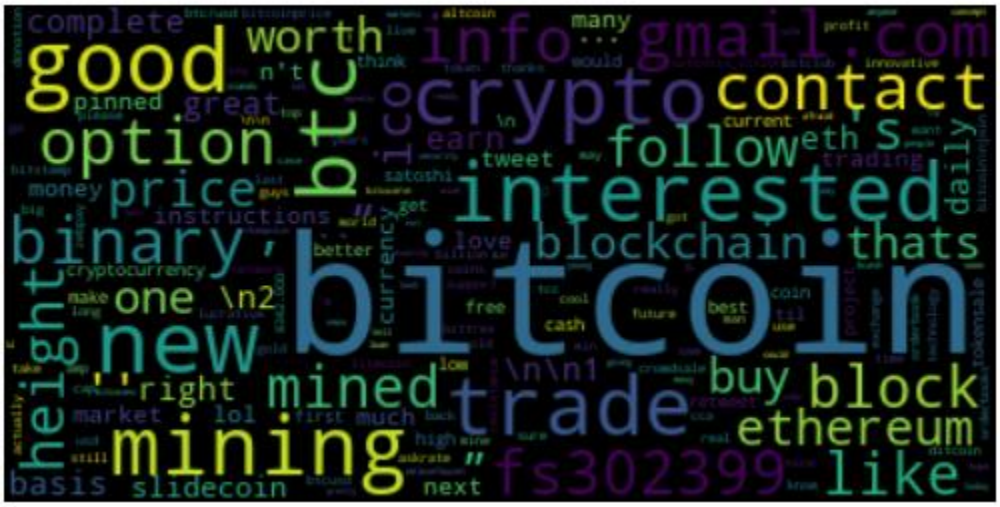

Figure 4 WordCloud for Sentiment

For instance, in the word cloud for positive sentiments, keywords such as 'good,' 'new,' and 'interested' appear prominently, indicating that tweets expressing positive sentiment often highlight new opportunities and positive developments in the Bitcoin space. Conversely, in the word cloud for negative sentiments, words like 'risk,' 'scam,' and 'loss' may be more prevalent, reflecting concerns and negative experiences related to Bitcoin. These visualizations help in quickly grasping the main themes and sentiments expressed in the tweets, and they serve as a useful tool for further qualitative analysis.

Beyond individual keywords, analyzing common phrases, such as bigrams (two-word combinations) and trigrams (three-word combinations), provides deeper insights into the context in which certain terms are used. This analysis involves identifying frequently occurring word pairs and triplets within the tweets, which can reveal more specific topics and sentiments. For example, the bigram analysis might highlight phrases like 'buy bitcoin,' 'bitcoin mining,' and 'crypto trade,' indicating prevalent activities and interests within the Bitcoin community.

By examining these common phrases and their associations with different sentiments, we can uncover patterns and themes that are not immediately evident from single-word analysis. For instance, positive sentiment tweets might frequently contain bigrams such as 'good investment' or 'great opportunity,' while negative sentiment tweets might include phrases like 'high risk' or 'avoid bitcoin.' This analysis helps to provide a more nuanced understanding of the discourse surrounding Bitcoin on Twitter and the specific contexts in which sentiments are expressed.

Cluster Analysis

The first step in our cluster analysis involved converting the textual data into numerical features, which is essential for applying machine learning algorithms. We utilized TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to transform the tweets into a matrix of numerical values. TF-IDF is a widely used technique in text mining that reflects the importance of a word in a document relative to a collection of documents. This method helps in emphasizing important words while reducing the weight of commonly used terms that may not carry significant meaning. The TF-IDF vectorizer was configured to consider the top 1,000 most relevant terms, resulting in a feature matrix where each row represents a tweet and each column represents a term's TF-IDF score.

Given the high dimensionality of the TF-IDF feature matrix, we applied Principal Component Analysis (PCA) for dimensionality reduction. PCA is a technique that transforms the original high-dimensional data into a lower-dimensional space while preserving as much variance as possible. This step is crucial for simplifying the data and making it more manageable for clustering. We reduced the dimensionality to two principal components, which allowed us to visualize the data in a two-dimensional space and facilitated more efficient clustering. The reduced feature set retained the most significant patterns in the data, ensuring that the clustering process would be both accurate and computationally efficient.

To determine the optimal number of clusters, we employed the elbow method as shown in figure 5, a heuristic used in determining the appropriate number of clusters in K-means clustering. This method involves running the K-means algorithm with different numbers of clusters and plotting the sum of squared distances (inertia) for each cluster count. The "elbow" point on the plot, where the inertia begins to decrease more slowly, indicates the optimal number of clusters. We performed this analysis for cluster counts ranging from 1 to 10 and observed the plot to identify the point of inflection. Based on the elbow method, we determined that the optimal number of clusters for our dataset was three.

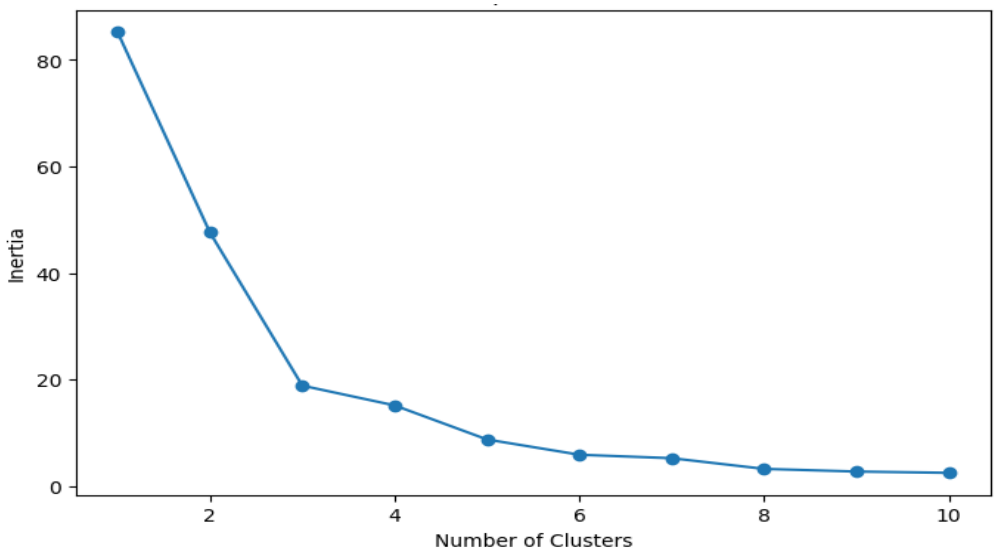

Figure 5 Elbow Method Plot

With the optimal number of clusters identified, we proceeded to perform K-means clustering on the reduced feature set. K-means is a popular clustering algorithm that partitions data into K distinct clusters based on feature similarity. Each tweet was assigned to one of the three clusters, with the algorithm iteratively refining the cluster centroids to minimize the variance within each cluster. The resulting clusters grouped tweets with similar characteristics, allowing us to identify distinct themes and patterns within the dataset.

The final step in our cluster analysis involved interpreting and labeling the clusters based on common themes or topics. We examined the tweets within each cluster to identify prevalent keywords and phrases, providing insights into the dominant themes. For instance, one cluster may have been characterized by discussions about Bitcoin trading and investment opportunities, indicated by frequent terms such as "trade," "investment," and "profit." Another cluster might have focused on concerns and negative experiences, with keywords like "risk," "loss," and "scam" being prominent. By analyzing the content of each cluster, we were able to label them accordingly and gain a deeper understanding of the various dimensions of sentiment and discourse surrounding Bitcoin on Twitter.

This comprehensive cluster analysis not only highlighted the diversity of topics discussed but also revealed underlying sentiment trends within the Bitcoin community. The insights gained from this analysis are valuable for identifying key areas of interest and concern, providing a foundation for further research and more targeted sentiment analysis in the future.

Result and Discussion

Sentiment Distribution Results

The sentiment distribution results from our analysis of Bitcoin-related tweets revealed a notable trend in the sentiment landscape. Out of the 1,544 unique tweets analyzed, the sentiment distribution was as follows: 829 tweets were categorized as neutral, 459 as positive, and 256 as negative. This distribution indicates a predominant neutral sentiment (53.7%), with positive sentiment accounting for 29.7% and negative sentiment comprising 16.6% of the total tweets.

This sentiment distribution suggests that the majority of the tweets about Bitcoin do not express strong emotions, which could be indicative of informational or factual content rather than opinionated posts. The significant proportion of neutral tweets implies that much of the discourse surrounding Bitcoin on Twitter involves sharing news, updates, or general information without an overt emotional tone. The positive sentiment, while smaller, still represents a substantial portion of the tweets, indicating a considerable level of optimism or favorable views towards Bitcoin. Conversely, the presence of negative sentiment, although the smallest, highlights that there are notable concerns and criticisms within the community.

The observed sentiment distribution has several implications for understanding the Bitcoin discourse on Twitter and its potential impact on market behavior. The predominance of neutral sentiment suggests that users often engage in sharing and consuming informational content, which could contribute to a more informed community. This type of engagement may help stabilize market perceptions, as factual and neutral information can counterbalance the effects

of extreme opinions and speculative sentiments. However, the substantial presence of positive sentiment indicates that there is a significant level of enthusiasm and optimism about Bitcoin, which can drive bullish market behavior. This optimism may be fueled by positive news, endorsements from influential figures, or favorable market conditions.

On the other hand, the presence of negative sentiment, although less frequent, cannot be overlooked. Negative tweets often reflect concerns about market volatility, security issues, regulatory challenges, or adverse news events. These sentiments can contribute to bearish market behavior and increased volatility as they might trigger fear, uncertainty, and doubt (FUD) among investors. Understanding the nuances of these sentiments is crucial for market analysts and traders, as it helps in anticipating potential market reactions and making informed decisions.

Overall, the sentiment distribution provides a comprehensive overview of the public's mood towards Bitcoin on Twitter. It highlights the diverse range of opinions and the factors that influence market sentiment. By continuously monitoring and analyzing these sentiments, stakeholders can better understand the underlying dynamics of the cryptocurrency market and develop strategies to navigate the complex and often volatile landscape. This analysis serves as a foundation for further research into the temporal dynamics of sentiment and its correlation with market movements, ultimately contributing to more effective sentiment-based trading strategies and risk management practices.

Keyword Analysis Results

The keyword analysis involved tokenizing the tweet text and performing frequency analysis to identify the most common words used in Bitcoin-related tweets. The top ten keywords in our dataset were 'bitcoin' (479 occurrences), 'new' (46 occurrences), 'good' (43 occurrences), 'crypto' (39 occurrences), 'trade' (39 occurrences), 'mining' (39 occurrences), 'btc' (38 occurrences), 'interested' (37 occurrences), 'gmail.com' (36 occurrences), and 'contact' (35 occurrences). These keywords provide insights into the predominant topics and themes discussed in relation to Bitcoin.

To visualize the frequency of these keywords, we created word clouds for each sentiment category: positive, neutral, and negative. Word clouds are graphical representations where the size of each word reflects its frequency in the text, providing an intuitive visual summary of the most prominent terms. In the positive sentiment word cloud, terms like 'good,' 'new,' and 'interested' appeared prominently, indicating that positive tweets often highlight opportunities, favorable developments, and engagement in the Bitcoin community. The neutral sentiment word cloud was dominated by general terms such as 'bitcoin,' 'crypto,' and 'trade,' reflecting a focus on information sharing and market activities without strong emotional tones. The negative sentiment word cloud featured terms like 'risk,' 'scam,' and 'loss,' suggesting that negative tweets are often centered around concerns and negative experiences related to Bitcoin.

The analysis of common keywords and phrases reveals distinct themes associated with each sentiment category. In positive sentiment tweets, the frequent use of words like 'good,' 'new,' and 'interested' suggests a focus on innovation, opportunities, and positive developments within the Bitcoin ecosystem. These tweets likely promote new projects, investment opportunities,

and positive market trends, contributing to a sense of optimism and enthusiasm among the community. Phrases like 'good investment' and 'great opportunity' were common in these tweets, emphasizing the positive outlook of users.

Neutral sentiment tweets, which comprised the majority of our dataset, were characterized by general and informational content. Keywords such as 'bitcoin,' 'crypto,' and 'trade' indicate that these tweets often discuss market activities, trading strategies, and general information about Bitcoin and other cryptocurrencies. The neutral tone of these tweets suggests that they are more focused on sharing knowledge, updates, and market analyses rather than expressing strong opinions. This category also included promotional content, as evidenced by the presence of terms like 'gmail.com' and 'contact,' indicating tweets that solicit engagement or advertise services.

Negative sentiment tweets, while fewer in number, highlighted significant concerns within the Bitcoin community. The frequent occurrence of words like 'risk,' 'scam,' and 'loss' points to the issues of market volatility, fraudulent activities, and financial losses. These tweets reflect the fears and doubts of users, often warning others about potential dangers or sharing negative experiences. Phrases such as 'high risk' and 'avoid bitcoin' were commonly found in negative sentiment tweets, indicating a cautionary tone aimed at preventing others from encountering similar problems.

Cluster Analysis Results

The cluster analysis of Bitcoin-related tweets was performed using K-means clustering on the TF-IDF vectorized data, reduced to two principal components for visualization purposes. The optimal number of clusters, determined using the elbow method, was found to be three. The K-means algorithm effectively grouped the tweets into these three clusters, with cluster sizes as follows: 1,346 tweets in Cluster 0, 163 tweets in Cluster 1, and 35 tweets in Cluster 2.

Visualizations of the clustering results, as shown in figure 6, were created by plotting the two principal components, with different colors representing each cluster. This visualization revealed distinct groupings of tweets, indicating that the clustering algorithm successfully identified meaningful patterns within the data. Cluster 0, being the largest, encompassed a broad range of tweets, while Clusters 1 and 2 represented more specific subsets of the data.

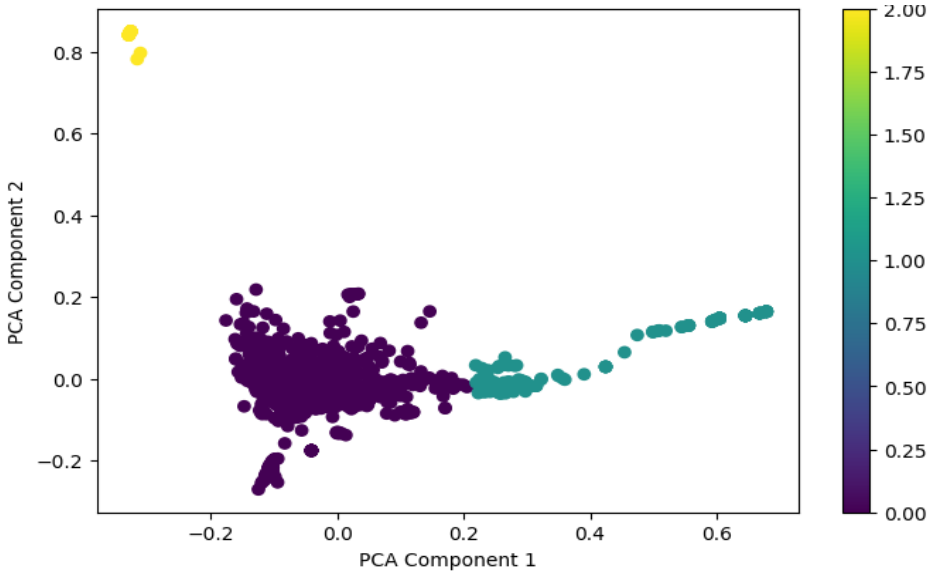

Figure 6 Cluster of Tweets

Upon examining the content of each cluster, we identified distinct themes and patterns that characterized the tweets within them. Cluster 0, the largest cluster, primarily consisted of neutral and informative tweets. Common keywords in this cluster included 'bitcoin,' 'crypto,' 'trade,' and 'news.' These tweets typically provided market updates, shared trading strategies, and discussed general information about Bitcoin without expressing strong emotions. The dominance of neutral sentiment in Cluster 0 suggests that this cluster represents the general discourse around Bitcoin, focusing on facts and information dissemination.

Cluster 1 was characterized by a higher concentration of positive sentiment tweets. Keywords frequently observed in this cluster included 'good,' 'new,' 'opportunity,' and 'investment.' Tweets in Cluster 1 often highlighted positive developments, investment opportunities, and favorable market conditions. The content of these tweets reflected an optimistic outlook towards Bitcoin, with users expressing enthusiasm about its potential and encouraging others to invest. This cluster can be seen as representing the optimistic and promotional side of the Bitcoin community, where users actively share positive news and endorsements.

Cluster 2, although the smallest, was notable for its concentration of negative sentiment tweets. Keywords such as 'risk,' 'scam,' 'loss,' and 'warning' were prevalent in this cluster. Tweets in Cluster 2 often expressed concerns about the risks associated with Bitcoin, including market volatility, scams, and financial losses. This cluster represented the cautious and skeptical viewpoint within the Bitcoin discourse, where users shared warnings and negative experiences to caution others. The presence of such a cluster highlights the significant concerns that exist within the community, despite the overall positive and neutral sentiment observed in the other clusters.

The clustering results provide a nuanced understanding of the Bitcoin-related discourse on Twitter. By categorizing tweets into distinct clusters based on sentiment and content, we can better comprehend the diverse perspectives and

themes that drive public opinion. This analysis not only enhances our understanding of the different dimensions of sentiment but also offers valuable insights for market analysts, investors, and communicators aiming to engage with the Bitcoin community. The identification of specific themes within each cluster can inform targeted communication strategies and help anticipate potential market reactions based on the prevailing sentiment.

Comparative Analysis

In comparing our findings with previous research studies on the influence of social media sentiment on cryptocurrency markets, several notable similarities and differences emerge. Our sentiment distribution analysis, which showed a predominant neutral sentiment followed by positive and negative sentiments, aligns with the results of earlier studies such as those by [13] and [14]. These studies also found that neutral and positive sentiments tend to dominate social media discussions about cryptocurrencies, with neutral tweets often conveying factual information and positive tweets expressing optimism about market trends.

Moreover, the clustering results from our study, which identified distinct groups of tweets characterized by neutral, positive, and negative sentiments, are consistent with the clustering patterns observed in prior research. For instance, [15] identified similar themes in their analysis of cryptocurrency discussions, noting clusters of tweets that focused on market updates, investment opportunities, and risk warnings. Our identification of these themes underscores the recurring topics and sentiments that pervade the Bitcoin discourse on Twitter.

However, there are also differences that highlight the novel contributions of our study. Unlike many previous studies that primarily focused on short-term sentiment fluctuations, our research provides a comprehensive, longitudinal analysis that captures sentiment trends over a more extended period. This approach allows for a deeper understanding of how sentiment evolves and the factors that drive long-term changes in public opinion about Bitcoin. Additionally, our detailed keyword and phrase analysis, which identified specific terms associated with each sentiment category, offers new insights into the language and topics that dominate different types of sentiment. This level of granularity in keyword analysis is less commonly addressed in earlier studies, providing a richer context for interpreting sentiment trends.

One of the key similarities between our study and previous research is the finding that social media sentiment, particularly on Twitter, significantly impacts cryptocurrency markets. This consistency reinforces the notion that Twitter is a vital platform for gauging market sentiment and predicting market movements. Our study's confirmation of these patterns adds robustness to the existing body of evidence, supporting the continued use of social media sentiment analysis in financial market research.

A notable difference in our study is the extensive use of clustering techniques to categorize tweets into thematic groups. While earlier studies have employed clustering, our application of K-means clustering combined with TF-IDF vectorization and PCA for dimensionality reduction provides a more refined and interpretable set of clusters. This methodological enhancement allows for a clearer identification of the dominant themes and their associated sentiments,

contributing to a more detailed understanding of the Bitcoin discourse.

Furthermore, our study's novel contributions include the integration of keyword and phrase analysis within the context of sentiment clusters. By examining the specific terms that characterize positive, neutral, and negative tweets, we offer deeper insights into the content and context of these sentiments. This approach not only enriches the understanding of sentiment dynamics but also provides practical implications for market analysts and investors who rely on keyword monitoring to gauge market sentiment.

Conclusion

In this study, we conducted a comprehensive analysis of Bitcoin-related tweets to understand sentiment distribution, keyword usage, and thematic clustering. Our sentiment distribution analysis revealed that neutral sentiments predominated, comprising 53.7% of the tweets, while positive sentiments accounted for 29.7% and negative sentiments made up 16.6%. This distribution highlights a balanced but predominantly neutral public perception of Bitcoin on Twitter.

The keyword analysis provided insights into the specific terms frequently associated with each sentiment category. Positive sentiment tweets commonly included words like 'good,' 'new,' and 'interested,' reflecting optimism and enthusiasm. Neutral tweets were dominated by general terms such as 'bitcoin,' 'crypto,' and 'trade,' indicating an informational focus. Negative sentiment tweets frequently mentioned 'risk,' 'scam,' and 'loss,' pointing to concerns and warnings within the Bitcoin community. The cluster analysis, utilizing K-means clustering, identified three distinct clusters, each characterized by specific themes and sentiment patterns, further enriching our understanding of the Bitcoin discourse on Twitter.

The findings from this study have several practical implications for investors, traders, and market analysts. By identifying the predominant sentiments and common themes in Bitcoin-related tweets, stakeholders can better gauge market mood and make more informed decisions. For instance, the high prevalence of neutral tweets suggests that a significant portion of the community is focused on factual information, which can stabilize market perceptions. Positive sentiment trends can signal potential bullish market behavior, while negative sentiments can serve as early warnings for market downturns.

The insights gained from keyword and cluster analysis can also enhance predictive modeling and market sentiment analysis. By integrating these findings into sentiment-based trading algorithms, analysts can improve the accuracy of market predictions and develop more effective trading strategies. Additionally, understanding the specific language and themes associated with different sentiments can help in crafting targeted communication strategies, whether to capitalize on positive trends or mitigate the impact of negative perceptions.

Despite the valuable insights provided by this study, there are several limitations to consider. The dataset, while extensive, may not capture the full spectrum of Bitcoin-related discourse on Twitter due to the use of specific keywords and the exclusion of non-English tweets. This limitation affects the generalizability of the findings to the broader global community. Additionally, the static nature of the

dataset does not account for the real-time dynamics of social media sentiment, which can fluctuate rapidly in response to news events and market changes.

Future research could address these limitations by incorporating a larger and more diverse dataset, including tweets in multiple languages and from various social media platforms. Additionally, employing real-time data collection and analysis methods could provide more immediate insights into sentiment trends and their impact on market behavior.

Building upon the findings of this study, future research can explore several promising directions. One potential area is real-time sentiment analysis, which would involve continuously monitoring and analyzing social media sentiment to provide up-to-date insights and predictions. This approach could leverage streaming data from Twitter and other platforms, combined with real-time processing capabilities, to enhance the responsiveness and accuracy of sentiment-based models.

Another important area for future exploration is the integration of sentiment analysis with other data sources, such as market transaction data, news articles, and regulatory announcements. This multimodal approach could provide a more comprehensive understanding of the factors influencing cryptocurrency markets. Furthermore, investigating the impact of sentiment expressed by influential accounts and key opinion leaders could yield deeper insights into the dynamics of market sentiment and its effects on Bitcoin prices. By addressing these areas, future studies can significantly advance the field of sentiment analysis in financial markets, offering more robust and actionable insights for market participants.

Declarations

Author Contributions

Conceptualization: T.W. and S.C.C.; Methodology: S.C.C.; Software: T.W.; Validation: T.W., S.C.C.; Formal Analysis: T.W., S.C.C.; Investigation: T.W.; Resources: S.C.C.; Data Curation: S.C.C.; Writing Original Draft Preparation: T.W. and S.C.C.; Writing Review and Editing: S.C.C. and T.W.; Visualization: T.W.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. C. McGregor, “Social Media as Public Opinion: How Journalists Use Social Media to Represent Public Opinion,” Journalism, vol. 20, no. 8, pp. 1070–1086, 2019, doi: 10.1177/1464884919845458.

[2] A. W. Prakasita, “The Use of Social Media (Twitter) in the Winning of East Java Province Election in 2018,” Indones. J. Polit. Stud. Ijps, vol. 2, no. 2, pp. 110–130, 2022, doi: 10.15642/ijps.2022.2.2.110-130.

[3] J. Du et al., “Use of Deep Learning to Analyze Social Media Discussions About the Human Papillomavirus Vaccine,” Jama Netw. Open, vol. 3, no. 11, p. e2022025, 2020, doi: 10.1001/jamanetworkopen.2020.22025.

[4] Q. Zhou and M. Jing, “Multidimensional Mining of Public Opinion in Emergency Events,” Electron. Libr., vol. 38, no. 3, pp. 545–560, 2020, doi: 10.1108/el-12-2019-0276.

[5] M. v. Klingeren, D. Trilling, and J. Moeller, “Public Opinion on Twitter? How Vote Choice and Arguments on Twitter Comply With Patterns in Survey Data, Evidence From the 2016 Ukraine Referendum in the Netherlands,” Acta Polit., vol. 56, no. 3, pp. 436–455, 2020, doi: 10.1057/s41269-020-00160-w.

[6] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, “Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning,” Entropy, vol. 21, no. 6, p. 589, 2019, doi: 10.3390/e21060589.

[7] S. Agarwal, S. Kumar, and U. Goel, “Social Media and the Stock Markets: An Emerging Market Perspective,” J. Bus. Econ. Manag., vol. 22, no. 6, pp. 1614–1632, 2021, doi: 10.3846/jbem.2021.15619.

[8] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011, doi: 10.1016/j.jocs.2010.12.007.

[9] L. Kristoufek, “BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era,” Sci. Rep., vol. 3, no. 1, p. 3415, Dec. 2013, doi: 10.1038/srep03415.

[10] D. Garcia, C. J. Tessone, P. Mavrodiev, and N. Perony, “The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy,” J. R. Soc. Interface, vol. 11, no. 99, p. 20140623, Oct. 2014, doi: 10.1098/rsif.2014.0623.

[11] J. Kaminski, “Nowcasting the Bitcoin Market with Twitter Signals.” arXiv, Jan. 18, 2016. doi: 10.48550/arXiv.1406.7577.

[12] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” Knowl.-Based Syst., vol. 89, pp. 14–46, Nov. 2015, doi: 10.1016/j.knosys.2015.06.015.

[13] N. Oliveira, P. Cortez, and N. Areal, “The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices,” Expert Syst. Appl., vol. 73, pp. 125–144, May 2017, doi: 10.1016/j.eswa.2016.12.036.

[14] R. C. Phillips and D. Gorse, “Predicting cryptocurrency price bubbles using social media data and epidemic modelling,” in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Nov. 2017, pp. 1–7. doi: 10.1109/SSCI.2017.8280809.

[15] L. Fang, E. Bouri, R. Gupta, and D. Roubaud, “Does global economic uncertainty matter for the volatility and hedging effectiveness of Bitcoin?,” Int. Rev. Financ. Anal., vol. 61, pp. 29–36, Jan. 2019, doi: 10.1016/j.irfa.2018.12.010.