Prophets and Profits: Twitter and the Future of Financial Forecasting

A THESIS

Presented to

The Faculty of the Department of Economics and Business

The Colorado College

In Partial Fulfilment of the Requirements for the Degree

Bachelor of Arts

By:

Dylan S. Brown

March 2024

Prophets and Profits: Twitter and the Future of Financial Forecasting

Dylan Brown

March 2024

Business, Economics and Society

Abstract

The rapid growth of social media platforms, particularly Twitter, has given rise to a new avenue of research exploring the intersection between online sentiment and financial market dynamics. This thesis investigates the relationship between the Twitter activity of Cathie Wood, a prominent figure in the financial industry, and the performance of her exchange-traded fund, ARK Innovation ETF (ARKK). By analyzing a comprehensive dataset spanning from August 2, 2021, to January 31, 2024, the study aims to uncover potential correlations between the frequency and sentiment of Wood's tweets and the volatility and returns of ARKK. The research employs regression analysis techniques, focusing on the beta of ARKK and the positive sentiment scores of Wood's tweets as dependent variables, while considering tweet frequency and ARKK's returns as independent variables. While the results did not yield statistically significant relationships between tweet sentiment and ARKK's performance, the study provides a valuable foundation for further exploration in the field of social media sentiment analysis and its application to financial markets. The findings contribute to the growing body of literature examining the intersection of social media and financial markets, highlighting the complexities and limitations inherent in such analyses. The thesis underscores the importance of considering a broader range of potential influencers and the need for creative approaches in future research endeavors.

KEYWORDS: (Twitter, Cathie Wood, ARKK, financial markets, social media, Sentiment)

ON MY HONOR, I HAVE NEITHER GIVEN NOR RECEIVED UNAUTHORIZED AID ON THIS THESIS

Signature: Dylan Brown

Table of Contents

Abstract .... **II**

1. Introduction **1**

2. Literature Review **4**

3. Data and Methodology **10**

4. Discussion and Conclusion **17**

References **22**

1. Introduction

The expeditious ascent of social media platforms has reshaped the landscape of communication, information dissemination, and data generation across various domains of society. Among these platforms, Twitter has emerged as a prominent force, boasting over 400 million active monthly participants (Statista). This immense user base has transformed Twitter into a potent tool for broadcasting messages to global audiences, while simultaneously generating vast public datasets rich with embedded signals pertaining to a wide array of topics. As financial markets and online behaviors continue to intertwine, the occurrence of financial discussions on Twitter has witnessed a remarkable surge.

Amidst the burgeoning relationship between Twitter and finance, fascinating possibilities arise surrounding the application of machine learning techniques to extract valuable insights from the platform's user-generated content. Specifically, the emotional states derived from the analysis of tweets, with millions of recurring online commentaries, hold promise in leveraging such information to inform decision-making within financial markets. Initially, straightforward methods of textual sentiment scoring appeared to be a promising avenue for the development of trading strategies. However, as scrutiny of these approaches has intensified, a far more intricate relationship between Twitter sentiment and price dynamics has been revealed, challenging the early reductionist assumptions.

This thesis aims to delve into the interplay between influential Twitter activity, particularly that of a prominent figure in the financial industry, and its potential to provide predictive advantages in the realm of equity markets. The research focuses on Cathie Wood, a renowned American investor, founder, and CEO of ARK Investment Management. Wood has garnered significant attention in recent years due to her successful investments in companies at the forefront of technological advancements, such as Tesla and Roku. Her exchange-traded fund, ARK Innovation ETF (ARKK), has experienced substantial growth since its inception in 2014, attracting a large following of investors who believe in her investment philosophy. (Lehtonen, 2024)

The primary objective of this study is to investigate the relationship between Cathie Wood's Twitter activity and the performance of her fund, ARKK. By analyzing a comprehensive dataset spanning nearly two and a half years, from August 2, 2021, to January 31, 2024, this research aims to uncover potential correlations between the frequency and sentiment of Wood's tweets and the volatility and returns of ARKK. The data collection process involved the use of the web scraping platform "Apify" to extract every tweet and retweet from Wood's Twitter account, along with their corresponding dates. Additionally, a sentiment analysis was conducted using Python to assign positive sentiment scores to each tweet based on the language used.

To gain a deeper understanding of the relationship between Wood's Twitter activity and ARKK's performance, this study employs regression analysis techniques. The initial regression focuses on the beta of ARKK as the dependent variable, aiming to explore its volatility at specific points in time. The independent variable in this analysis is the number of tweets posted by Cathie Wood per day. By examining the correlation between tweet frequency and ARKK's beta, this study seeks to uncover potential patterns and insights.

Furthermore, a second regression is conducted to expand the scope of the analysis and consider a broader range of potential influencers. In this regression, the dependent variable is the positive sentiment score associated with each of Cathie Wood's tweets, while the independent variable is the returns data for ARKK. By aligning the dates of the daily returns with the corresponding positive sentiment scores, this analysis aims to investigate the impact of tweet sentiment on the fund's performance.

The findings of this study contribute to the growing body of literature examining the intersection of social media and financial markets. By focusing on a prominent figure like Cathie Wood and her influential presence on Twitter, this research provides valuable insights into the potential of leveraging social media data for predictive purposes in the realm of finance and equity markets. Moreover, the study highlights the importance of considering the limitations and complexities inherent in such analyses, emphasizing the need for further research and refinement of methodologies.

The subsequent sections of this thesis will delve into the existing literature on the topic, outlining key developments and debates surrounding the analysis of tweets for predictive equity pricing models. The methodology employed in this study will be discussed in detail, including the data collection process, sentiment analysis techniques, and regression models used. The results of the analyses will be presented, along with a comprehensive discussion of their implications and potential limitations. Finally, the conclusion will summarize the key findings, highlight the contributions of this research to the field, and propose avenues for future investigation.

2. Literature Review

2.1 Overview

Social media has fundamentally transformed communication capacities, information velocity and data generation volumes across society. The social media platform Twitter with over 400 million active monthly participants has enabled rapid broadcasting of messages to global audiences, concurrently producing vast public datasets with embedded signals pertaining to a myriad of topics (Iqbal, 2024). The occurrence of financial discussions revolving around the aforementioned have increased immensely as markets and online behaviors continuously intertwine.

Amidst the growing relationship between Twitter and finance, tantalizing possibilities arise surrounding the harnessing of machine learning to mine value from expressions. Specifically, emotional states calculated with such mined values through platforms like Twitter, with millions of recurring online commentaries indicate promise in utilizing such information to take action within financial markets. Initially, simple textual sentiment scoring methods appeared to be promising for trading strategies. But rising scrutiny reveals a far more intricate relationship between Twitter and actual price dynamics challenging early reductionist assumptions.

This literature review examines key developments and debates within the domain of analyzing tweets to develop predictive equity pricing models. It examines prevalent Twitter data harvesting tactics, processing techniques that have been adopted, uncovers performance results, dissects analytical horizons, discusses ethical considerations, and reviews potential revelations of financial behavioral foresight.

2.2 Approaches to Twitter Data Sourcing

The foundational step that supports overall analytic veracity involves the identification of relevant tweets for collection. Researchers often target messages by accounts with professional finance connections based on follower thresholds or textual content suggesting expertise (Nofer et al., 2013). For instance, Arora (2019) filtered millions of tweets by key terms that indicate financial knowledge. This focused sampling isolates the valuable commentary that could possibly shift asset values and investor actions. However, others argue that incorporating broader sentiment tracking that considers non-professional users as well enhances model comprehensiveness. The inclusion of both professional consensuses along with the noise of the greater public mood revolving around the markets could paint a more accurate picture (Gunter et al., 2012). Data harvesting tactics remain debatable, but creative alternatives show promise in encapsulating relevant commentary. Snyder (2019) bypassed manual filtering entirely by training unsupervised networks to recognize predictive tweet features regardless of author. Zhang (2013) recommends focusing on keyword search that dynamically responds to correlation. Gunter (2012) notes accelerating discussion volume can also provide analytical signals beyond obvious sentiment.

The way in which sourcing decisions mature as web architectures simultaneously evolve and accessibility to historical messages improves will unveil powerful indications for the potential of data analysis for such purposes. But defining exactly which communication branches offer the most validity remains unclear presently, though arguably where strongest revelations can be located is developing.

2.3 Sentiment Scoring and Alternative Processing Avenues

Assuming possession of a valuable Twitter dataset, prevalent data processing tends to fixate on textual sentiment analysis, accomplished by algorithmically evaluating positive-negative aversion of word choices as the dominant predictive agent. Drus (2019) demonstrates lexicon methods that generate aggregate emotional valence scores from tweets scalably. However, Drus (2019) also highlights lexicon's limitations relating to semantic complexities, showing how the training of neural networks on tweets that target context-specific expressions can improve accuracy notably.

Beyond obvious textual signals, there are alternative analytical dimensions that leverage Twitter's data richness to be explored (Kearney et al., 2013). Elements including discussion volumes, hashtag choices and author influence show potential for financial relevancy, though often discarded during classical focused text scoring.

Pursuing such can form new comprehensions of future potential but requires the hurdling of data handling and modeling prior. Though rationales seem abundant as proof accumulates, useful inferences remain unturned. Powerful sculpting that amalgamates layers of data and their correlation to finances hold promise in unlocking behavioral knowledge that classical assumptions lack.

2.4 Relationships Between Twitter Analytics and Market Behaviors

A popular narrative presumes there to be readily discoverable predictability between online sentiment among vocal financial influencers and actual investment actions that follow. Under such assumptions, baseline time series analysis that test lead-lag daily correlations between Twitter mood composite scores present as a justifiable inaugural position prior to the commencement of fancier assessments (Kearney et al., 2013). The assumption that there exists daily viral passions or panic's that are digitally expressed would indicate overlapping of feasible trading behaviors that are measurable - this is plausible.

However, modern financial markets exhibit resilience, arbitraging simplistic signals faster than models assume. Alternate evaluation approaches have gained credence by embracing such inherent complexities. Focused event analyses surrounding extreme return outliers finds Twitter data assists in understanding the origins of unusual volatility bursts via related commentary more reliably than generalized performance prediction (Umar et al., 2021).

In sum, evolving perspectives recognize that skillful integration of Twitter analytics alongside classical indicators can offer a matured path forward. Rather than solo forecasting silver bullets, social data measures are capable of providing supplementary signal detectors, explanatory enhancements, and confidence buffers for strategy. But exactly which asset classes and time frames social datasets inform most proficiently remains an open situation.

2.5 Wider Perspectives on Twitter’s Analytic Horizons

While the dominant investigations concentrate on seemingly obvious message text sentiment quantification, wider analytical horizons remain nearly untapped outside rudimentary gestures (Gunter et al., 2012). Immense stores of years’ worth of archived behavioral signatures persist largely unexplored by most finance focused studies with the scale of such daunting classical handling.

Yet rationales feel insistent to progress in uncovering the untapped terrain, where useful inferences await in using metadata to contextualize message exchanges. Enhanced tooling can ease extractions from such to dissect sources and their interconnectedness to unlock richer economic information.

Inevitable developments that enable accessible synthesizing of Twitter's embedded data holds promise for unlocking the usage of one-to-many communication for financial purposes. Those that endure the growing pains of the evolving tactic could reap the rewards of speech becoming an infrastructure that challenges the status-quos.

2.6 Ethical Debates on Financialized Twitter Analytics

Rapidly advancing analytic capacities inevitably raise ethical considerations regarding consent, privacy, manipulation, and social impacts from financially incentivizing large-scale public data mining (Royakkers et al., 2004). Twitter enables democratized everyday communication, seemingly without oversight. The repurposing of trails that users generate for the purpose of predicting their collective economic activities risks the financial weaponization of individual words upon the collective.

However, public benefit through such holds promise as well if technologies harnessed balance creativity against evil intent and unintended damages, but unrestrained applications risk private interests wielding inequitable power that lacks accountability. Though there exists techniques that democratize analytical leverage, appropriate governance appears crucial given the financial sphere's historic appetite and willingness to undermine organic cultural spaces to fulfill it. Ethical guidelines must coincide with implementation, enforcing those leveraging Twitter analysis to incorporate the collective rather than marginalize upon their vulnerability.

2.7 Conclusion

This literature review explored key developments involving the analysis of influential Twitter activity and its collective sentiment states capability for providing market predictive advantages. It reveals mixed evidence surrounding popular textual polarity approach’s ability to provide market timing edges and invites the questioning of the qualifications that such claims hold. Select studies show potential value in longer term Twitter mood scoring with traditional indicators and integrative design.

Meanwhile the technique of democratization to combat ethical dilemmas around consent, data privacy and social impacts demands urgent address. With weaponization potential just as real as the tool’s potential benefits, those exploring Twitter’s unlocking for investment forecasting must pursue progress while condemning mal usage. If, when and how social data can and will add value in predicting markets at acceptable trade-offs is up for debate. But with analytical frontiers expanding faster than governance thus far, the future infrastructure will inevitably see restructuring to balance the relations between speech, economics, and rights.

3. Data and Methodology

3.1 Overview

Cathie Wood is a prominent American investor, founder, and CEO of ARK Investment Management, an investment management firm. Wood gained significant attention in recent years due to her successful investments in companies at the forefront of technological advancements, such as Tesla and Roku (Lehtonen, 2024). Her exchange-traded fund (ETF), ARK Innovation ETF (ARKK), has experienced substantial growth since its inception in 2014, attracting a large following of investors who believe in her investment philosophy.

The research that I conducted took into account every one of Cathie Wood's Tweets from August 2, 2021, through January 31, 2024 - a span of nearly two and a half years. The platform "Apify" is an online custom web scraper that provides its users with the ability to scrape any website for data to be extracted. Through this site, I was able to extract every tweet and retweet from Wood's twitter account, with the corresponding dates. With such data, I was able to run a simple sentiment analysis via Python that produced Positive sentiment scores for each tweet, based solely on the wording of said tweets.

I also obtained the daily market data for Wood's Fund (ARKK). With the date of every tweet, further the number of tweets per day, the sentiment score of each tweet, and the overlapping returns from the ETF (ARKK), I was then able to define my dependent and independent variables and conduct my regressions.

3.2 Dependent Variable

In my initial regression analysis, I focused on the beta of the ETF (ARKK) as the dependent variable, aiming to gain insights into its volatility at specific points in time.

While numerous statistics are associated with the performance of stocks and ETFs, beta provides a more focused perspective on volatility. With that said,

3.3 Independent Variable

The independent Variable I used for the initial regression was the number of tweets tweeted per day by Cathie Wood. I was able to calculate this simply by creating an excel function that counts tweets with the same date correlated to them. Some days, Wood never tweeted, and others she tweeted as many as six times. With this data, I sought to investigate whether there exists a correlation between the beta of ARKK and the frequency of Cathie Wood's tweets.

3.4 Findings:

Linear Regression: Model 1

Linear regression

beta	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Nooftweets	.004	.023	0.19	.849	-.041	.05	
Constant	2.105	.044	47.62	0	2.018	2.192	***

Mean dependent var	2.113	SD dependent var	0.369	
R-squared	0.000	Number of obs	265	
F-test	0.036	Prob > F	0.849	
Akaike crit. (AIC)	226.950	Bayesian crit. (BIC)	234.109	

**** p<.01, ** p<.05, * p<.1*

Upon examining the regression model that utilizes the number of tweets as a predictor variable for the beta coefficient, several key insights emerge. First and foremost, the model's R-squared value is notably low, indicating that the number of tweets alone accounts for a minimal portion of the variation in the predicted beta values. This suggests that there are likely other significant factors influencing the beta coefficient that are not captured by this model.

Furthermore, the t-value associated with the number of tweets variable falls below the critical threshold of 1.96, implying that the relationship between the number of tweets and the beta coefficient is not statistically significant at the conventional 95% confidence level. In other words, we cannot reject the null hypothesis that there is no genuine relationship between these two variables based on the available evidence.

Moreover, the p-value corresponding to the number of tweets variable is 0.849, which exceeds the typical significance levels of 0.01, 0.05, and 0.10. This further reinforces the notion that the number of tweets does not have a statistically significant impact on the beta coefficient, even at the more lenient 90% confidence level. Consequently, we cannot conclude that the number of tweets is a reliable predictor of the beta coefficient based on this regression analysis.

The dependent variable of beta in my regression analysis posed certain challenges in establishing a significant relationship with the number of tweets by Cathie Wood. Beta, a measure of ARKK's volatility relative to the overall market, is computed using a relatively long window of data, making it sticky and resistant to sudden changes. In other words, beta values do not fluctuate drastically in short timeframes. Given the limited duration of this study, spanning from August 2, 2021, to January 31, 2024, it is not surprising that there was a struggle to find a significant correlation between tweet volume and beta.

Summary Statistics: Model 2

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
beta	265	2.113	.369	1.305	2.766
Nooftweets	265	1.638	.983	1	6
PositiveSentiment	265	.475	.016	.408	.518
ret	265	-.002	.03	-.088	.118

One can see in the summary statistics of beta that the standard deviation of .369, along with the other measures of mean, min, and max display an overall lack of volatility due to the very nature of the metric itself.

For a significant relationship to emerge, Cathie Wood's tweeting patterns would need to exhibit a specific trend. For instance, if Wood consistently concentrated her tweets during periods of high market volatility and remained silent for most of the other times, there might be a possibility of observing a correlation. However, even in such a scenario, the limited time window of this analysis reduces the likelihood of capturing substantial changes in beta.

The choice of beta as a dependent variable, coupled with the relatively short study period, posed challenges in establishing a significant relationship with the volume of Cathie Wood's tweets. Beta's inherent stability over short timeframes makes it less responsive to the influence of tweet volume within the confines of this analysis.

Acknowledging the limitations of this model is crucial. To refine our understanding of the factors driving the beta coefficient, alternative approaches may assist in developing a more accurate predictive model. By expanding the scope of my analysis and considering a broader range of potential influencers, a more nuanced understanding of the complex dynamics at play in determining the beta coefficient would be possible.

Ultimately, while the current regression model provides valuable insights into the relationship between the number of tweets and the beta coefficient, it is crucial to recognize its limitations and pursue further to refine our understanding of this important financial metric.

3.5 Dependent Variable II:

For my second regression, with the goal of expanding the scope of my analysis and considering a broader range of potential influencers, my dependent variable was the positive sentiment score correlated to each of Cathie Woods tweets. Sentiment scoring assigns a numeric value to the positive and or negative tonality of, for the purpose of this study, Cathie Woods tweets. Via python, I was able to assign each of

Woods tweets a Positive sentiment score that ranges between 0.00 (the most negative possible) to 1.00 (the most positive possible). All the positive sentiment scores fall between .40 and .60, with the vast majority falling below .50, meaning her tweets tend to lean negative rather than greater than .50 which would be positive leaning.

Figure 1

Python Code

```
from transformers import BertTokenizer, BertForSequenceClassification
import torch
import pandas as pd

# Load pre-trained BERT model and tokenizer
model_name = 'bert-base-uncased'
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertForSequenceClassification.from_pretrained(model_name)
```

Figure 2

Python Code

```
df = pd.read_csv("text.csv")
df.columns.tolist()
texts = df['full_text'].tolist()

positive_scores = []
negative_scores = []
i=0

final_dict = {}

for text in texts:
    print(i)
    # Tokenize the text
    tokens = tokenizer.encode_plus(text, return_tensors='pt', padding=True, truncation=True)

    # Perform inference
    with torch.no_grad():
        outputs = model(**tokens)

    # Get predicted probabilities
    probs = torch.softmax(outputs.logits, dim=1).tolist()[0]

    # Print the sentiment score
    positive_score = probs[1]
    negative_score = probs[0]

    i+=1

    final_dict[i] = {'positive_score': positive_score, 'negative_score': negative_score}
```

3.6 Independent Variable II:

The independent variable that I utilized for my second regression was the returns data for ARKK. This data is public information and was organized to have the dates of the daily returns aligned with the dates of the daily positive sentiment scores that I calculated.

3.7. Findings II:

Linear Regression: Model 3

Linear regression

ret	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
PositiveSentiment	-.102	.116	-0.88	.38	-.331	.127	
Constant	.047	.055	0.84	.4	-.062	.155	

Mean dependent var	-0.002	SD dependent var	0.030
R-squared	0.003	Number of obs	265
F-test	0.772	Prob > F	0.380
Akaike crit. (AIC)	-1096.160	Bayesian crit. (BIC)	-1089.001

*** *p*<.01, ** *p*<.05, * *p*<.1

We sought to examine the potential impact of positive sentiment on the returns of Cathie Woods fund ARKK. To do so, we refined our dataset by excluding observations with no tweets on the respective days, resulting in the removal of 364 observations. This was done so that positive sentiment scores of zero assigned to days with zero tweets were not included.

It is essential to note certain limitations inherent in our study. Firstly, positive sentiment might not serve as a robust signal, as it could relate to various aspects, some of which may be irrelevant or even negative concerning Cathie Wood's firm. This nuance adds complexity to the interpretation of the relationship between positive sentiment and ARKK’s returns.

The study's timeframe, spanning from 2021 to 2024, may be relatively short, potentially impacting the robustness of our findings. Gathering additional data over an extended period could enhance the study's depth and provide a greater understanding of the dynamics at play.

The analysis yielded non-significant results for key metrics. The R-squared value, indicative of the model's explanatory power, did not reach significance. Similarly, the t-value associated with positive sentiment failed to surpass the critical threshold, suggesting a lack of statistical significance. Furthermore, the p-value, used to assess the significance of the relationship, exceeded conventional levels.

Focusing on contemporaneous returns fails to capture any delayed or gradual effects of tweet sentiment on the ETF's performance. Using raw returns instead of abnormal returns also hinders the model's ability to isolate the specific impact of tweet sentiment, as it doesn't control for overall market performance. These limitations likely contributed to the lack of a significant relationship observed between tweet sentiment and ARKK's returns.

Given these non-significant findings, it becomes evident that our model did not identify a statistically significant impact of positive sentiment on the dependent variable. These results highlight the importance of considering the study's limitations and suggest potential avenues for further research. To gain a deeper understanding of this potential impact, future research should explore incorporating time lags, utilizing abnormal returns, and considering alternative sentiment measures.

One possible extension could involve a more nuanced exploration of sentiment, considering specific contexts or employing advanced sentiment analysis techniques. Additionally, the suggestion to conduct an event study could offer insights into the impact of specific events on the observed relationships within the dataset. While our current analysis provides a foundation, future research should explore these avenues to refine our understanding of the intricate dynamics surrounding positive sentiment and its potential impact on the chosen dependent variable such as returns.

4. Discussion and Conclusion

The rapid evolution of social media platforms, particularly Twitter, has given rise to a fascinating realm of research exploring the intersection between online sentiment and financial market dynamics. This thesis delved into the complexities between the Twitter activity of a prominent figure in the financial industry, Cathie Wood, and the performance of her exchange-traded fund, ARK Innovation ETF (ARKK). By analysing a comprehensive dataset spanning nearly two and a half years, this study aimed to uncover potential correlations between the frequency and sentiment of Wood's tweets and the volatility and returns of ARKK.

The findings of this research contribute to the growing body of literature examining the potential of leveraging social media data for predictive purposes in the realm of finance and equity markets.

The initial regression analysis, which focused on the relationship between the number of tweets posted by Cathie Wood per day and the beta of ARKK, revealed several key insights. Although the model's explanatory power, as indicated by the R-squared value, was notably low, it highlighted the presence of other significant factors influencing the beta coefficient that were not captured by the tweet frequency alone. Furthermore, the lack of statistical significance in the relationship between tweet volume and ARKK's beta, as evidenced by the t-value and p-value, suggests that the number of tweets may not serve as a reliable predictor of the fund's volatility.

These findings underscore the importance of considering a broader range of potential influencers when examining the complex dynamics at play in the financial markets. While the frequency of Cathie Wood's tweets alone may not provide a comprehensive picture, the study's second regression analysis investigated the role of sentiment in predicting ARKK's returns. By focusing on the positive sentiment scores associated with each tweet and aligning them with the corresponding daily returns, this analysis sought to uncover the impact of tweet sentiment on the fund's performance. It is crucial to acknowledge the limitations of this study. The positive sentiment scores, derived from a sentiment analysis of Wood's tweets, may not serve as a robust signal in isolation. Sentiment expressed in tweets can relate to various aspects, some of which irrelevant concerning the performance of ARKK. This nuance highlights the complexity of interpreting the relationship between sentiment and financial outcomes, as the sentiment expressed may not always directly translate to the intended impact on the fund's returns.

Moreover, the relatively short timeframe of the study, spanning from 2021 to 2024, may limit the robustness of the findings. Financial markets are subject to a wide range of influences, both short-term and long-term, and capturing the full extent of these dynamics may require an extended period of analysis. Future research could benefit from gathering additional data over a longer timeframe to enhance the depth and reliability of the insights obtained.

Despite the limitations, this study provides a valuable foundation for further exploration in the field of social media sentiment analysis and its application to financial markets. The non-significant results obtained in the second regression analysis, as indicated by the R-squared value, t-value, and p-value, suggest that the positive sentiment scores alone may not have a statistically significant impact on ARKK's returns. However, these findings should not discourage future research endeavour’s, rather, they highlight the need for a more nuanced approach to sentiment analysis and the consideration of additional factors that may influence the relationship between sentiment and financial outcomes.

One potential avenue for future research could involve a more granular examination of sentiment, considering specific contexts and employing advanced sentiment analysis techniques. By moving beyond the binary classification of positive and negative sentiment and considering the intensity, context, and subject matter of the tweets, researchers may uncover more meaningful insights into the relationship between sentiment and financial market dynamics. Additionally, conducting event studies that focus on the impact of specific events or announcements on the observed relationships could provide a more targeted understanding of how sentiment influences market reactions in real-time. Furthermore, future studies could expand the scope of analysis to include a broader range of social media platforms and influential figures within the financial industry. By examining the sentiment expressed across multiple platforms and by a diverse set of market participants, researchers can gain a more comprehensive understanding of the collective sentiment landscape and its potential impact on financial markets. This approach could also help identify any platform-specific biases or differences in sentiment expression, allowing for a more nuanced interpretation of the findings.

Another important consideration for future research is the incorporation of additional financial metrics and market indicators alongside social media sentiment analysis. By integrating sentiment data with traditional financial data, such as trading volume, market capitalization, and macroeconomic indicators, researchers can develop more robust models that capture the complex interplay between sentiment and market dynamics. This holistic approach could potentially improve the predictive power of the models and provide a more comprehensive understanding of the factors driving market movements.

As the field of social media sentiment analysis continues to evolve, it is crucial to address the ethical considerations surrounding the use of such data for financial gain. The privacy concerns associated with mining public sentiment data and the potential for manipulation or misuse of this information warrant careful examination and the development of appropriate guidelines and regulations. Researchers and practitioners alike must navigate these ethical challenges purposefully to ensure that the insights obtained from social media sentiment analysis are used responsibly and in a manner that benefits society as a collective.

In conclusion, this thesis has explored the intersection between social media sentiment, specifically focusing on the Twitter activity of Cathie Wood and the performance of ARKK. While the findings may not have yielded statistically significant results in terms of predicting ARKK's returns based on sentiment alone, they have laid the groundwork for further investigation and highlighted the complexities involved in this area of research. The study's limitations, such as the relatively short timeframe and the need for more nuanced sentiment analysis techniques, provide valuable insights for future research endeavors.

As we stand at the precipice of a new era, where the boundaries between the digital and the financial worlds continue to blur, the potential for harnessing the impact of social media sentiment in the realm of finance is immense. By embracing the challenges and opportunities presented by this evolving landscape, researchers and practitioners can unlock new frontiers of understanding, and drive innovation in the field of financial market analysis. The insights gained from studies like this will not only shape the future of investment strategies but also contribute to the broader discourse on the societal impact of social media and its role in shaping one’s economic reality. The true value of this research lies not in the immediate findings, but in the avenues it opens for further exploration. With every development in understanding how social media posts can affect financial markets on a microeconomic level, we move closer to a future where the power of collective sentiment can be harnessed to create a more informed, transparent, and equitable financial world for all.

References

"Apify Console." *Apify*, console.apify.com/actors/heLL6fUofdPgRXZie/runs/6d6z3siBWLFoAoMHg#output. Accessed 11 Mar. 2024.

Arora, Anuja, et al. "Measuring Social Media Influencer Index- Insights from Facebook, Twitter and Instagram." *Journal of Retailing and Consumer Services*, vol. 49, no. 49, July 2019, pp. 86–101.

Cano-Marin, Enrique, et al. "Twitter as a Predictive System: A Systematic Literature Review." *Journal of Business Research*, vol. 157, Mar. 2023, p. 113561, https://doi.org/10.1016/j.jbusres.2022.113561.

Daily, Investor's Business. "5 Cathie Wood Stocks to Watch Include This EV Giant." *Investor's Business Daily*, 4 Mar. 2024, www.investors.com/research/best-cathie-wood-stocks-to-watch-coinbase-block-roku-zoom-video-tesla-falls/. Accessed 11 Mar. 2024.

Daniel, Kent D., et al. "Investor Psychology in Capital Markets: Evidence and Policy Implications." *SSRN Electronic Journal*, 2001, https://doi.org/10.2139/ssrn.278848.

Dixon, S. "Topic: Twitter." *Statista*, 21 Mar. 2023, www.statista.com/topics/737/twitter/#topicOverview.

Drus, Zulfadzli, and Haliyana Khalid. "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review." *Procedia Computer Science*, vol. 161, no. 1, 2019, pp. 707–714, reader.elsevier.com/reader/sd/pii/S187705091931885X?token=407B16D221E8BAFC0601AC0CE5183C3177DBF8A7918F6C5D06F7C9F60D7AF0A267F3E00AAB27BEE74EC9D4C5D0D25202, https://doi.org/10.1016/j.procs.2019.11.174.

Fiala, Vojtěch, et al. “Impact of Social Media on the Stock Market: Evidence from Tweets.” *European Journal of Business Science and Technology*, vol. 1, no. 1, 11 Nov. 2015, p. 24, https://doi.org/10.11118/ejobsat.v1i1.35.

Gunter, Barrie, et al. *Sentiment Analysis a Market-Relevant and Reliable Measure of Public Feeling?*

Karami, Amir, et al. “Twitter and Research: A Systematic Literature Review through Text Mining.” *IEEE Access*, vol. 8, 2020, pp. 1–1, https://doi.org/10.1109/access.2020.2983656.

Kearney, Colm, and Sha Liu. “Textual Sentiment Analysis in Finance: A Survey of Methods and Models.” *SSRN Electronic Journal*, 2013, https://doi.org/10.2139/ssrn.2213801.

Nisar, Tahir M., and Man Yeung. “Twitter as a Tool for Forecasting Stock Market Movements: A Short-Window Event Study.” *The Journal of Finance and Data Science*, vol. 4, no. 2, June 2018, pp. 101–119, https://doi.org/10.1016/j.jfds.2017.11.002.

Nofer, Michael, and Oliver Hinz. *Using Twitter to Predict the Stock Market Where Is the Mood Effect?* 2015.

Ranco, Gabriele, et al. “The Effects of Twitter Sentiment on Stock Price Returns.” *PLOS ONE*, vol. 10, no. 9, 21 Sept. 2015, p. e0138441, https://doi.org/10.1371/journal.pone.0138441.

Rao, Tushar, and Saket Srivastava. *Analyzing Stock Market Movements Using Twitter Sentiment Analysis*. 2012.

Snyder, Luke , et al. *Interactive Learning for Identifying Relevant Tweets to Support Real-Time Situational Awareness*.

"Twitter Revenue and Usage Statistics (2024)." *Business of Apps*, www.businessofapps.com/data/twitter-statistics/#:~:text=data%2C%20The%20Information-.

Umar, Zaghum, et al. "A Tale of Company Fundamentals vs Sentiment Driven Pricing: The Case of GameStop." *Journal of Behavioral and Experimental Finance*, vol. 30, no. 30, June 2021, p. 100501, https://doi.org/10.1016/j.jbef.2021.100501.

van Wel, Lita, and Lambèr Royakkers. "Ethical Issues in Web Data Mining." *Ethics and Information Technology*, vol. 6, no. 2, 2004, pp. 129–140, https://doi.org/10.1023/b:etin.0000047476.05912.3d.

Zhang, Lei. *Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation*. 1 Jan. 2013.