CORP3400 Strategic Management Dissertation

An Investigation into The Effects of New Media on The Financial Markets and its Implications for the Efficient Market Hypothesis

Case Study of GME short squeeze

Arron Wilkins – P2534355

Word count: 8,332

Date: 20/04/2023

Supervisor: Jack Brown

Abstract

This research uses the case study of the GameStop (GME) short squeeze of 2021, to investigate the influence new media has had, specifically Reddit, on financial markets and the implications this has for the Efficient Market Hypothesis (EMH). A mixed-methods approach is utilised, together with Reddit post data analysed using the VADER sentiment model, with the financial variable from FRED and Yahoo finance. A Vector Autoregression (VAR) model is employed to examine and quantify the relationships between these variables.

The findings explore the gaps within the literature and reveal correlations between new media sentiment and financial market trends, indicating that platforms such as Reddit can influence trader behaviour, effectively challenging EMH. This study contributes to the understanding the dynamics between new media and financial markets, offering valuable insights for future research and policymakers.

KEY WORDS:

New Media, Reddit, Subreddit, WallStreetBets, Financial Markets, GameStop, VADER Sentiment, VAR model.

Contents

Chapter 1 – Introduction

1.0 Background

In an ever-evolving world, the rapid increase in new media platforms such as social media and online forums/blogs (Brown, D., 2019), has vastly transformed the way in which information is consumed and communicated. Information, trends, opinions, and insights that were once guarded secrets and exclusive to financial professionals are commonplace on such platforms, empowering the individual traders and investors. Logically, this has sparked an interest in a search for an understanding into the potential impact of the new media's impact on financial markets and implications with the rational traders of the efficient market hypothesis (EMH) (Fama, E., 1970).

EMH stands to reason that if the traders operating within financial markets are rational, therefor so should the markets they trade within be rational, making it impossible for a truly efficient market to yield abnormal returns through trading strategies (Fama, E., 1970). This emergence of new media platforms, more specifically, Twitter and Reddit, have created a unique contributor to the widespread of information in all forms, this potentially influencing the financial markets.

A significant event that brought major worldwide attention is the GameStop (GME) Short Squeeze of 2021. This involved users of the subreddit r/WallStreetBets collectively engaging retail trades to drive the price of the GME stock, this all-in to combat the market expectations and causing losses for large hedge funds holding short positions in the stock. Notable event that will be researched for many years to come, as it demonstrated the potential power of new media and bring disruptions to traditional markets whilst challenging the assumptions of EMH.

1.1 Research Aim and Objectives

The aim of this research is to determine the effects of new media on financial market, focusing on the potential for traders to engage in high-risk strategies breaking the EMH with significant influence form new media. The objectives of this study are:

- Investigate the relationship between new media Sentiment and financial market variables.
- Analyse the impact new media is having on the financial stock prices.
- Examine the implication of the findings for EMH.

1.2 Research Significance and Rationale

This research is significant as it seeks to identify the role played by new media within the financial markets, and its implications for the assumptions of EMH. Through interpreting the sentiment of the new media, this study can provide insight into the ways in which these platforms are amassing large numbers of traders that can challenge traditional assumptions of

markets. The findings of this research could prove to be valuable to academics, policymakers, or other researchers in the field of finance, as they offer an alternative perspective on the evolving relationship of new media and financial markets.

Utilising the case of the GameStop short squeeze, and other similar events, the growing importance of new media in altering market dynamic can be evidenced, leading to the rationale of this research. Exploring the sentiment of the new media and financial trends and patterns, this study aims to find, if any, the impact, and potential consequences resulting from GME short squeeze on financial markets.

1.3 Research Methodology overview

To address the prior section, research aims and objectives, this research will make use of a mixed-method approach, employing both qualitative data and quantitative data. Qualitative data will be gathered from Reddit, in post forms, this data will then be subjected to Valence Aware Dictionary for Sentiment Reasoning (VADER) sentiment analysis (Beri, A., 2020) to identify trends and patterns in the sentiments. Quantitative data will be gathered from financial source, the Federal Reserve Economic Data (FRED) and Yahoo Finance, collected to access the trends in the market and potential impact of new media on the trends. Vector autoregression (VAR) model (Korstanje, J., 2021) is applied to forecast and interpret the relations between the variables under investigation.

1.4 Structure

The research will be laid into 6 key identifiable chapters. Flowing from the introduction into the literature review, leading the reader through the steps taken to identify gaps in which to conduct the 3rd chapter. The methodology details the data collection and the VADER and VAR models. Chapter 4 bring the findings and presents them with analysis, shedding light to key trends and patterns identified within the Reddit and financial variables. The fifth chapter covers the implications of the results, and potential reasonings for any discrepancies observed. Finally, the sixth chapter is the conclusion, where a summary of the main findings and discussion of the limitation of the dissertation can be found, furthermore future research and recommendation to the implications and contributions that are brought about through this research.

Chapter 2 – Literature Review

2.1 INTRODUCTION

Over 1.7 billion monthly visits (Statista, 2022) Reddit is a form of new media that has greatly impacted the way financial markets (Hurst, A. 2022) are interacted by investors. With the world constantly evolving, it is only natural that how people consume media and information changes along with it (Adgate, B., 2021). New media is defined as "new information and entertainment products and services that use digital technologies such as the internet" by the Oxford Dictionary (OLD, 2022), however "New media" is a very broad term, for this research we must shorten it to specific applications such as Twitter, Reddit and YouTube. These are some of the main forms of new media, more accurately these are the forms that will be mentioned in this paper and how they are impacting the investor of the modern day. A narrowed look into the financial crisis event known as the 'GameStop short squeeze' in 2021 (Bloomberg, 2021), and how Reddit had an influence into the investor's behaviours and overall outcome of the company stock performance. An important distinction is that Reddit is comprised of various forums and networks, being referred to as a "Subreddit", which can delve into interests and or hobbies. The main subreddit discussed will be WallStreetBets (Dictionary, 2018).

The structure of this literature review will be comprised of THREE key areas: *Theories* on Consumers and Pricing, *Empirical Review* of the research and finally the *Gaps* between the theorems and reality. Focusing on the Theories and framework of Aggregate Demand-Aggregate Supply, Behavioural Financing, Investor Consumer Behaviour, and how media creates shock factors which affected the way consumers invest.

2.2 THEORY

2.2.1 The Efficient Market Hypothesis

A great base for this review is to define the efficient market hypothesis [EMH] (CFI, 2022) which brings about the conclusion that stocks are traded at the most suitable market value, as they are representative of all available public information about the traded stock. There are critiques of this theorem, one notably that investors can overreact to information, creating an inefficient market environment, which arguably is what happened with Bitcoin during May of 2021 (BBC, 2021, CNBC, 2021) Elon Musk makes a tweet about banning the use of the currency. EMH also relies on the assumption that everyone holds the same personal valuations for a stock, and all agrees on the fair price, Fichan, R. and Rhodes, P. (2005) discuss how individuals are differing in fundamental ways, therefore with a competitive economy way of predicting an individual's personality could become increasingly more important to predict the economy. EMH is a vital point to start this discussion due to r/WallStreetBets [WBS] being a modern quintessential critique of the EMH theory as this was all publicly available knowledge that had sparked an overreaction into the stock market.

2.2.2 Demand-Supply framework

In the book "Money, Banking and Financial Markets" CECCHETTI, S.G. and SCHOENHOLTZ, K.L. (2017) layout the framework for the Aggregate Supply and Demand. Aggregate demand is comprised of four main factors, (CFI, 2022).

1. Investment Spending	**2. Consumption Spending**
3. Government Spending	*4. Net Exports*

For the scope of this research, we do not have the word count nor time to take all four factors into account with demand features. For this we will focus on the impacts that media has on influencing the Consumption Spending and in part the Investment Spending.

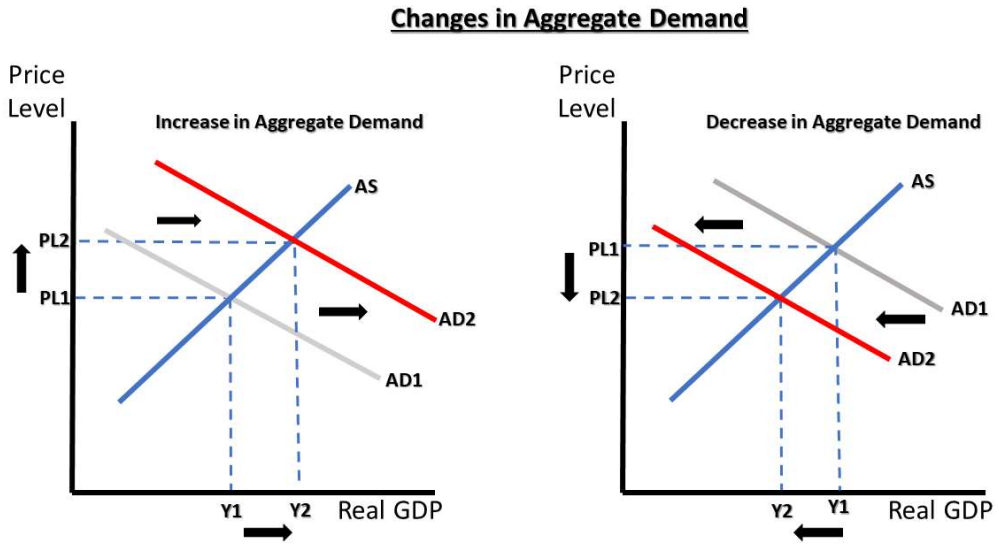


Figure 2.2.2

It is an important to understand that Cecchetti and Schoenholtz (2017) show that the impact of change in the aggregate demand can have short- and long-term effects for the economy. Notably the short run effects of a right shift to aggregate demand, outputs and increase to the price and GDP, which are factors that impact the economy greatly. If it is believed that one source of this change can arise from 'Consumer Confidence' and 'Business optimism' (CECCHETTI et al., 2017) then we must investigate whether media can impact the confidence of consumers and the optimism of businesses, and these factors can have major impact on demand curves and pricing.

2.2.3 Behavioural Finance and Consumer Behaviour Theory

Finding a metric or framework that is reliable and consistent in measuring how we behave is a seemingly impossible task, no singular framework or model is the best nor perfect, the way a person behaves changes constantly. Martin, P. and Bateson, P. (2007) elaborate on this topic well in the book; "Measuring Behaviour" summarises that although it is useful to have a metric for behaviour, it is a very difficult task to undertake any possibilities as outliers.

In the Investor Behavior, Baker et al., (2014) states that predicting the behaviour of investors is becoming an increasingly more sort after form of research into predicting and understanding the decisions that investors make. Mention of various heuristics for investors, one example is Animal spirits (Farmer, R.E.A. 2010; Akerlof, G.A. and Shiller, R.J. 2009) which is a discussion on human irrational emotions having effect on investment rather than the intrinsic value of stock. This is disputed to be noise (De Long et al., 1990) that the theory of EMH (CFI, 2022) accounts for, instead of ignored, the research shows the relevance of irrational investing could be on the rise.

2.3 EMPIRICAL

2.3.1 Foundational Research

Over many decades there has been a fair amount research on the effects, intended or not, that media has on the financial markets, which is reflected in the share prices of companies, Berner,(1979) explores the effects that the news outlets have on the price of a company's stock during and after releasing news on the company, even though this research is dated and unable to reflect the scope of the effects of reddit and twitter have on the stock performance. The researcher deems it to be relevant to understand that before new media had been established news publications had major effects in financial markets and were not considered as just noise in the stock market. Critiques with this research soul focuses into Oil companies. Furthermore, Berner only investigates the change of the prices, never delving into the aspects of why the investors react when these faced with what the media releases, only share prices changes.

With the early iterations of forums being introduced, Wysocki (1999) finds correlations between posts made on stock forums of Yahoo! Messaging boards and abnormal behaviour in the pricing of stocks, stating that a prediction could be made with the volume of posts during out of hour trading and the volume of trades and stock returns during the next day. This is the earliest iteration of research into 'new media'. However, the research had limited amount of data, with it only looking into top 50 companies, the effects of volume change could had been stated to not occur from these forums alone. An argument could be made that, even with news for these larger firms being publicly available these lesser-known forums would be a place for people to build biases and share similar information which could be argued would have an impact of their investing as a whole.

2.3.2 Modern Research

More recently, Tetlock, (2007) evaluates the possibility of a strategy that uses the pessimism output form the media to predict the pricing of stock. Tetlock separated the equation with differing variable, one example is two equations, one with prolonged window of returns, and another with a much more concise window for returns. The research stated the conclusion to be that high values of media pessimism are a causation for decreases to market prices. This is limited with the research neglecting to account for the possibility of momentum trading, being the factor that affects the cost as a result after the release of the media content.

Research into new media within investments, which have been analysing relations between social media and company stock prices. Arthur,J, C (2013), investigated the relation of social media accounts on Facebook and a company's stock performance. The research indicated a relationship between the price of a brand's stock and the quantity of Facebook accounts made around the stock by the consumers, the popularity of a company's social media presence had a significant correlation to the performance of a company's stock price. However, the research is limited in the method of the collection of those Facebook accounts and providing evidence that these accounts are anything other than an already existing correlation of popularity, rather than being a causation for popularity which it tries to suggest.

2.3.3 After GME 2021

This literature review will now extend into the literature that came after the crisis event of the GME short squeeze during 2021, the notability and power of change, in how people viewed the power of new media, is represented by the vast increase into research around these types of communities. Furthermore, the research following these events draws upon distinct separation between research before and after this crisis event.

WallStreetBets is found to be a major factor for the pricing of shorted stocks. Song, (2021) finds correlation between the number of discussions on the forum and the future prices of stocks. Song's tested varied sets of regressions. The initial regression finds that stocks with short squeeze experience and greatly affected by the discussions on WallStreetBets (Song, J., 2021). The second and third regressions find discussion with purposeful short squeeze goals provides evidence for their effect on stock price to be much greater than taking discussions that show no aim. This research lends a position to investigate the effects of WallStreetBets effect on intraday trading within taking hourly price data into account based upon discussions with aim. Limitations exist within the small scale of stocks considered; these limitations are mentioned within the research. However as mentioned in another piece of research, Diangson, B., Jung, N., (2021) mentions limitations around the influx of media attention that this forum had on the financial markets. Therefore, the popularity created substantial noise in the data.

Further papers discuss the increase in the irrational trading strategy that developed from the WallStreetBets forum, Gendron, Y., et.al., (2022) concludes it to be organised chaos, however it is noted that the new development of digital financial platforms may favour the 'trial and error' approach of these irrational investors. It is also noted that this literature states small individual investors can use new media to construct themselves into larger entities that make similar trades. Witts, et al. (2021) provides further research into the irrational trading that

has developed from the new media forums. Limited with the data gathering methods nor being able to encapsulate every form of modern discussion with the original creator's intension, example is shorthand financial terms or even memes that related in video form or images.

Interestingly, Fedyk, V. (2022) takes the research to the financial applications that are being adopted by the new form of irrational investors. Fedyk finds relevance of activity on WallStreetBets and investors making use of Robinhood's application for their investments into short positions, finding an indication of a notable network existing for such high volumes of individual retail traders aligning on similar high-risk positions. Although this paper neglects to mention the events after GME showing WallStreetBets seeking new platforms for trading due to the actions of Robinhood during the crisis which arguably halted the potential profits of the short squeeze.

2.4 GAP IN THE LITERATURE

The literature shows a gap in the effective speed of new media on financial markets. The literature has limited insight into the shorter time-periods of markets, research shows they neglect the speed in which the new media can affect markets, with the ability to transfer actions and ideas in much faster fashions, the literature mentioned do not establish the current potential of this newly developed irrational investment strategies. Research shows there is a gap in literature that explores the research of new media empowering traders to be more high-risk in lower time frames and break the efficient market hypothesis. Furthermore, the literature alludes to a potential investigation into a behavioural bias being formed within these forums that stores "consumer confidence" within markets that these individual traders would not normally consider. The literature does not access the significant increase in size of retail investors and creation of accounts into these retail trading applications with correlation in new media.

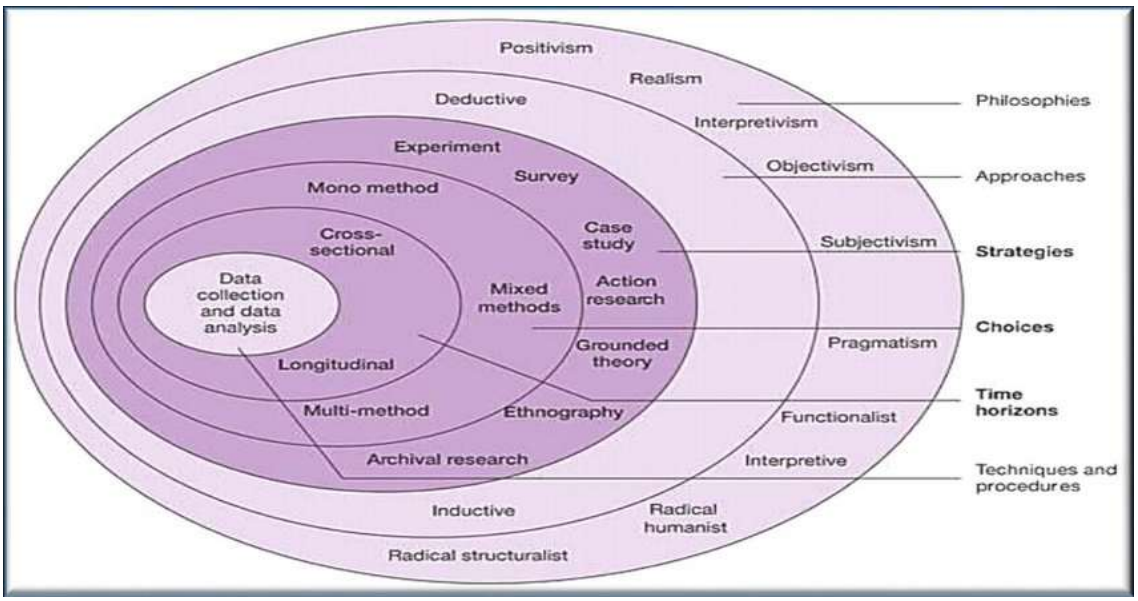
2.5 CONCLUSION

Traditional forms of media showed similar effects on markets, with companies issuing reports and articles, etc. which would create a reaction in the market within accordance with EMH. However, with the scale of traders, being accessed by many more individuals, these reactions are on such a greater scale over a much shorter period of time, without the factual evidence these new media sources become a limitation to the efficient market hypothesis.

Chapter 3 – Methodology

3.1 Introduction

This chapter's aim is to outline and justify the research method employed to investigate the 'Effects of New Media on Financial Markets' and to what extent new media influences traders to take on high-risk behaviours in lower time frames, potentially breaking the efficient market hypothesis (EMH) (Fama, E., 1970). In this section the research philosophy, approach and strategy used in the study will be discussed, followed by a detailed insight into data collection and analysis techniques. By employing a rigorous and appreciate methodology, the research seeks to provide new information into the impact new media is having on trading behaviour and financial market dynamics.



The research onion (Saunders et al., 2015)

3.2 Research Philosophy

For this study, pragmatic research philosophy is used, as Saunders et al. (2015) recognised the value of both objective and subjective knowledge and allows for the use of multiple methods to address the research question. Pragmatism is applicable to the research's aim of investigating the effects of new media on financial markets, as it enables the integration of quantitative data from FRED and yahoo, with the qualitative insight of new media platforms, such as Reddit, through VADER sentiment analysis (Saunders et al., 2015).

By making use of the pragmatic philosophy, this research can freely depict the complex relationship between new media and financial market behaviour, benefiting from the strengths of positivist and interpretivist approaches (Saunders et al., 2015). The pragmatic perspective gives flexibility within the research process, concentrating on the optimal methods required to achieve a thorough comprehension of the impact new media has on financial markets.

Pragmatism provides a strong framework for the multi-method research strategy and inductive approach employed in this study (Saunders et al., 2015).

3.3 Research Approach

Inductive approach for this research is deemed most suitable, for exploring the research aims. An inductive approach delves into the unknown relationship and patterns in data and then derive conclusions or new theories from these themes, Saunders et al., (2009) writes 'theory follows data rather than vice versa'.

A pragmatic philosophy and an inductive approach, allows the research to identify patterns and trends in the data which can capture the complex relationship between new media and financial market behaviour, providing a thorough insight of the research problem. As mentioned by Diangson et al., (2021) there is limited research or new theory on this topic, inductive approach enables researchers to newly generate such hypothesis and theories based upon observed data (Saunders et al., 2015).

3.4 Research Choice

To investigate the new media's effects on financial markets, a multi-method choice is used. As Saunders et al, Lewis and Thornhill (2015, p166) explain, multi-method research provides a "scope for a richer approach to data collection, analysis and interpretation". Including quantitative and qualitative data collection and analysis methods the research gives a greater understanding into the complexity of the relationship of the research problem.

Both aspects of data, qualitative and quantitative, in this study will offer insights into the new media's role in shaping the decisions traders make through VADER sentiment, and the evidence of this impact through the changes within the pricing of the chosen stock to identify, in this case it is the GameStop (GME) stock price. These datasets can then be analysed together, with further financial variables, to uncover a comprehensive understanding of the research issue. Multi-method aligns well with pragmatism, and inductive approach of the research to open multiple perspectives and types of data to create conclusions.

3.5 Research Strategy

For research strategy, this study chooses the case study approach, examining the GameStop (GME) short squeeze of 2021, a highly significant event that impacted the financial world, that was greatly influenced through by new media (Song,,J.,2021), the Reddit subforum: r/WallStreetBets. Focusing into this case study in detail, a sentiment analysis conducted into the posts within the subreddit, specific key words that drove the research were terms such as "GME", "GameStop", "GAMESTOP". These were used to filter the posts collected in the research's qualitative data, run in the VADER sentiment analysis. Using this filter, we can eliminate non-essential posts that do not relate to the case study event in the research, and better gauge depth of insight with the relationship of new media and the financial market data.

Case study comes with strengths, namely the ability to give an extensive understanding of the specific environment in which the research problem arises. This enable both quantitative and qualitative data investigation which aligns well with the research multi-method choice. The case study approach does, however, come with limitations and potential biases. A key limitation is the potential lack of the case study being generalised (Simon, M.K. and Goes, J., 2013), as the findings of one study may not apply to all accounts of the situation or contexts. Furthermore, the analysis of prior events could introduce hindsight bias into the study, as the researcher might be inclined to concentrate on data that support the report results.

Regardless of these limitations, the case study can provide an excellent base to build future research upon, tackling the complex interplay of new media's role in the financial markets in a real-world scenario (Saunders et al., 2015)

3.6 Research Data

3.6.1 Scope and Sauces

To conduct research into new media and the effects on financial market, limitations must be created, the data sources Reddit API, FRED API, and Yahoo Finance, can give access to historical Reddit posts, Unemployment Rate, Consumer Price Index, S&P500 data. The rationale behind these sauces, are the credibility of the FRED data, and the base line that is created from testing these particular financial market variables. S&P500 was chosen over other indices based on prior research in this area (Arthur J, C 2013) (Levine, M., 2021), and is a great comparator for stocks. The use of secondary data, Consumer Price Index (CPI) and Unemployment rate (Unrate), give the credibility to the results from testing and comparing these variables together (Saunders et al., 2015).

The Data is limited to the dates from January 2021 to April 2021, this is due to limitations surrounding computer capacity and time restraints, faced by the researcher, processing, and analysing large amounts of data would yield more accurate results, however, requires greater hardware and amounts of time necessary.

3.6.2 Data Collecting and Difficulties

Conducting the collection of data can be procured through various methods, one such method is Python. Python gives way to the use of premade packages that will decrease the potential errors and therefor increase the repeatability of the process. To gather data from Reddit, the script needs to requestion OAuth token which requires a Reddit Application Programming Interface (API) to securely communicate the programming to the Reddit application (Briggs, J., 2020). With this enabled, the data can be retrieved in raw form which is not ready for the sentiment analysis in this state. Financial Data from the Federal Reserve Economic Data (FRED) also requires API access to retrieve data and can be extracted from a specified date.

Difficulties encountered during the data collection, complications with the GME and S&P500 data require these datasets to be retrieved from Yahoo, a straightforward and reliable source of information, code for this is found in the Appendix. Furthermore, delays with client and

server-side response times limit the amount of data able to be obtained, this can be overcome by using another API such as pushshift API, or the 'pmaw' package in python to optimise these issues (Briggs, J., 2020). The Sentiment data also then had to be converted into a time series dataset format, this however can bring difficulties along with it, such difficulties are how you go about processing the grouping of mentioned compound score data into a time frame. There issues encountered on this matter lead the scope of the research to be narrowed due to lengthy coding problems which began to test the researcher's abilities with python errors more so than the interpreting the research at hand.

3.6.3 Data Cleaning and Pre-processing

This section is vital for the process of data collection, various other research pieces on sentimental analysis have included these steps, (Abraham et al., 2018) (Kwon et al., 2019), it prepares the raw datasets to be in a format better suited for modelling and interpreting. For the sentimental analysis, the dataset was cleaned by:

- **Combining the 'title' and 'body'**: A single 'text' column to ensure the data captures all relevant information for the sentiment analysis.
- **Text normalisation**: Converting the text into lowercase and removing special characters, numbers and extra spacing. Ensures the words with similar meaning are treated as the same.
- **Date Format**: Converting the timestamp column into a datetime object in a format usable in time series analysis. Issues arose that made the code unusable do to faults in handling the sentiment data into time series.

Additional steps, such as tokenization, stopword removal, stemming or lemmatization, proved to not be necessary steps in the case of a VADER sentiment analysis library is designed to handle this kind of raw text data internally. For the financial data the steps taken for cleaning were:

- **Load and Merge**: Merging the datasets into a singular DataFrame (DF) and aligning by date.
- **Handling Missing Data**: By making use of the fillNa function, the missing data values can be filled using forward fill (fill NaNs with most recent non-NaN value)
- **First Differencing**: Applying first differencing on all the financial datasets to make them stationary, a crucial step in VAR modelling, Korstanje, J. (2021) states the VAR

model can only work "if each variable in the model is stationary".

- **Scaling**: scaling the 'Sentiment' column by 100 to make it more comparable with other variables in the dataset.

- **Lag Selection**: Akaike Information Criterion (AIC) to determine an optimal number of Lags for the VAR model.

Difficulties with removing NaNs came up within the code, and potential infinity issues within the dataset arose, however these were solved through package loading issues and not issues with the data.

3.6.4 Sentiment Analysis and VAR model

Prior research around this topic (Critien et al., 2022) conducts a similar study that analysed the twitter sentiment through VADER and showed that the VADER model was applicable for this task and more notably used amongst researchers of this field. The VAR model allows the testing of correlation between multiple time series variables (Korstanje, J. 2021), as the primary data used in this research is time series, comparing all of these under the VAR model is a very practical and logical next step for interpreting the datasets.

A VADER sentiment analysis allows researchers to understand the opinions and trend of the posts on Reddit. As stated by Hutto et al., (2014), it is 'quick and computationally economical without sacrificing accuracy'. It holds limitations when working with longer or more complex text, however these posts are much more suitable for the VADER model as it works well with slang, emoticons, and other informal elements, which are highly present on Reddit (Song, J., 2021).

Since the data type used in this research is time series, a VAR model is an effective choice for simultaneously handling multiple timer series variables (Korstanje, J. 2021). The research encountered issues with the VAR model forecasting in python, this was due to limited background in python coding. Limitations of the VAR model in this research are in the assumptions of the model. The VAR model assume a linear relationship between variables, which, in the real-world scenarios is not guaranteed (Korstanje, J. 2021). A major strength of VAR model is the flexibility, in that the research can use this to forecast results (Korstanje, J. 2021).

Storing and organising the data was essential to keeping efficiency within the code, especially for repeatability and credibility. The Reddit data collected was stored in a '.csv' file format, with the columns of raw data into date, title, body, this was a limitation of the data collection, as more information on the data could have allowed for further analysis such as number of likes and comments, this data would have allowed to filter the data with interactions and weight posts accordingly.

3.7 Ethical Considerations

This research approach used of a case study, in collecting historical posts from Reddit and analysing in conjunction with publicly available financial data, raises no ethical concerns. As the Reddit API is open-sourced and all information is freely and publicly available to be used in research.

Chapter 4 – Findings and Analysis

4.1 Introduction

This chapter of the research brings insight into attempting to answer the research aim, of how new media effects financial markets, through a case study of GameStop Short Squeeze 2021. The research gathered several thousands of posts from r/WallStreetBets during the period of, 2021, January 28th to April 1st to use during the VADER sentiment analysis. The Findings chapter will display the financial data and the potential correlations of the datasets with the sentiment analysis. The final part will present the findings from the VAR model and conclude with a summary of these key findings. During the research there proceeds to be various instances within the coding process that alludes to the data failing VAR model assumptions and not showing promising signs for effective predictive probabilities.

4.2 Sentiment Analysis Findings and Interpretation

Through the timeline of GME mentions within Reddit, the data can be allocated to represent the total mentions without a gauge of sentiment. The graph below show mentions of GME during two separated periods, notably 2021 has a significant larger number of mentions, making the potential impact greater on the markets during this time, however for the scope of the research only gathers sentiment of GME mentions in after Jan 28th 2021.

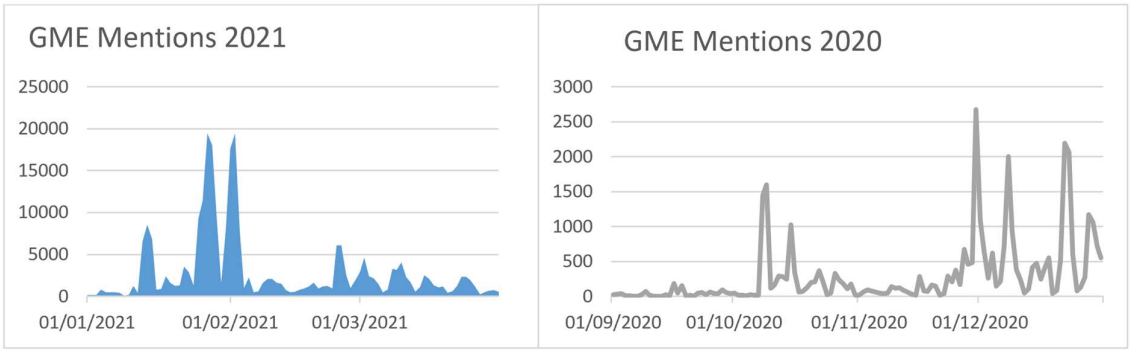


Fig. 4.2.1

VADER sentiment can be interpreted through the polarity score it provides for given text, this ranges from negative to positive, giving the researcher insight into the outcome of textual data based on search parameters, for this case GME and GameStop. VADER produces three component scores, Positive (pos), Negative (neg), and Neutral (neu), then a compound score (compound), which represents the overall sentiment of the text, -1 being most negative to +1 being most positive.

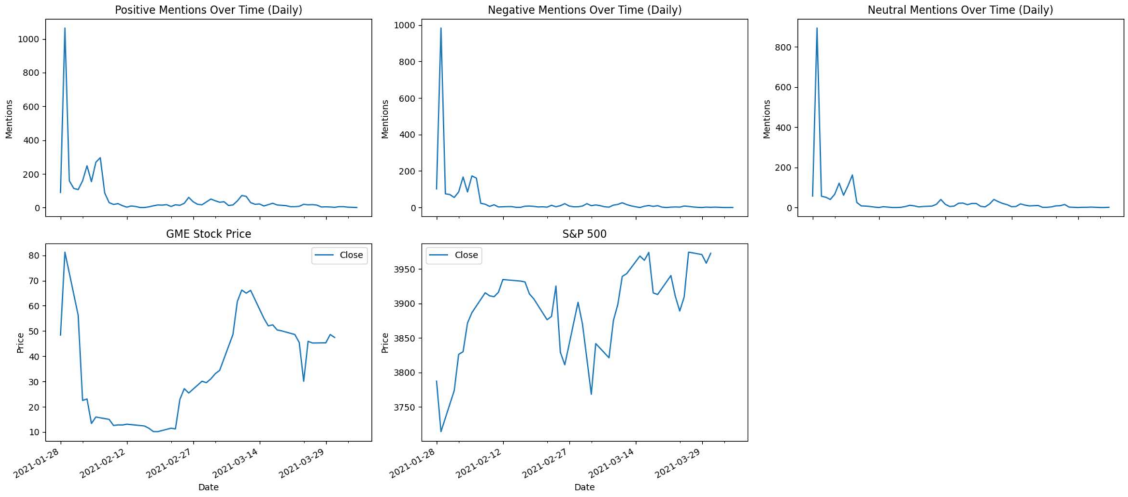


Fig. 4.2.3

These graphs Represent the Sentiment data mentions along with the GME stock pricing and S&P 500, we can see from this that the sentiment mentions, along the top, appear to be closely linked to one another. This could be due to a trend within the data that we are not accounting or identifying for, for example Diangson, B. and Jung, N. (2021) mentioned this may be in due effect to the nature of the posts on the subreddit, as they mention their positions taken, this increases mentions, yet the VADER model might struggle discerning these correctly. A noticeable trend in the data is the sentiment mentions decreasing significantly after January's end.

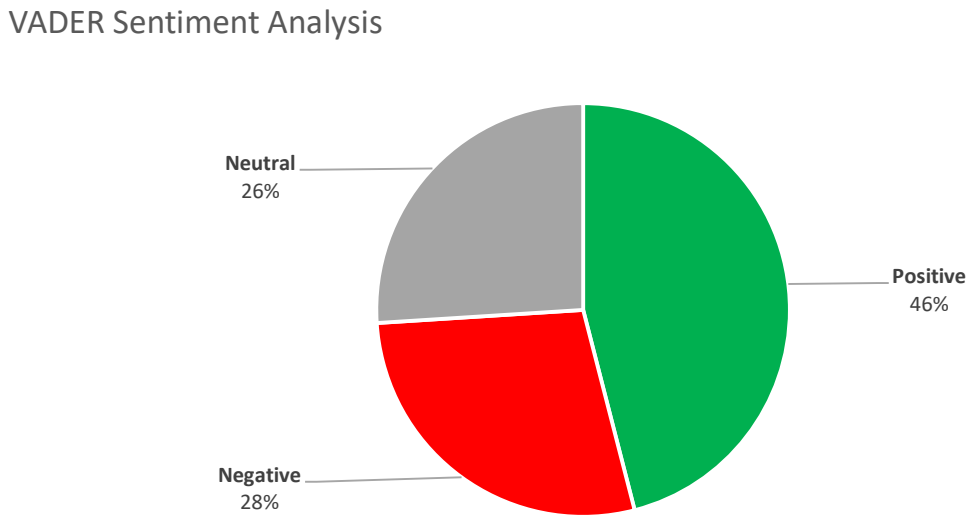


Fig. 4.2.4

Inferring from the graph in Figure 4.2, the data is majority positive sentiment towards "GME", with an almost equal negative and neutral remaining sentiment. These percentages can reflect the intention and trend of the subreddit's opinion on the GME stock, hence showing a significant push in a positive direction towards the stock based on a majority positive output from VADER analysis. Yet this does not mean these results are definitive and guaranteed to show causation for the financial data changes.

4.3 VAR Model

```
Number of NaN values in the dataset:
Positive    0
Negative    0
Neutral     0
GME         0
SP500       0
CPIAUCSL    0
UNRATE      0
dtype: int64
```

```
Number of inf values in the dataset:
Positive    0
Negative    0
Neutral     0
GME         0
SP500       0
CPIAUCSL    0
UNRATE      0
dtype: int64
```

Fig. 4.3.1

Fig. 4.2.5 shows, the combined data of the financial and sentiment data, key pre-processing step, ensuring there are no NaNs or Infinite values, if there are Infs the code will replace them as NaNs and fill the data forward (Korstanje, J. 2021).

```
Positive - ADF Statistic: -6.6959138849083715
Positive - p-value: 4.0001870983888e-09
Negative - ADF Statistic: -21.05719828472153
Negative - p-value: 0.0
Neutral - ADF Statistic: -2.699965368509103
Neutral - p-value: 0.07404426199163917
GME - ADF Statistic: -1.6218894950767626
GME - p-value: 0.471738847325829
SP500 - ADF Statistic: -2.105512912162601
SP500 - p-value: 0.24230361800429623
CPIAUCSL - ADF Statistic: -0.053801255614717755
CPIAUCSL - p-value: 0.9538500803022787
UNRATE - ADF Statistic: -1.046580050736635
UNRATE - p-value: 0.7358986659857341
Neutral_diff - ADF Statistic: -12.476520800933343
Neutral_diff - p-value: 3.15944683708857e-23
GME_diff - ADF Statistic: -7.545617114579743
GME_diff - p-value: 3.2896554653060514e-11
SP500_diff - ADF Statistic: -6.110393745390707
SP500_diff - p-value: 9.363306627673162e-08
CPIAUCSL_diff - ADF Statistic: -8.247529007065113
CPIAUCSL_diff - p-value: 5.502487591661047e-13
UNRATE_diff - ADF Statistic: -8.124038404635957
UNRATE_diff - p-value: 1.1357657574188362e-12
```

Fig. 4.3.2

This is the Augmented Dickey-Fuller (ADF) (Cheung, et al., 1995) test, which is a statistical test used in econometrics to test timeseries data for stationarity. The ADF Statistic represents the likely hood of stationarity, the more negative value the higher likelihood of stationarity within the data set. P-value in the ADF indicates whether to accept of reject the Null Hypothesis H_0 = the time series is non-stationary). If the p-value < 0.05 we can reject the null hypothesis, meaning the time series is stationary, if p-value > 0.05 we cannot reject the null hypothesis and consider the time series non-stationary (Korstanje, J. 2021).

The Result shown here lead to Fig. 4.2.6 right test with the '_diff' where the data has now gone through first order difference, ensuring stationarity through a secondary ADF test.

4.3.1 Optimal Lag in the VAR model

```
Results for equation GME
================================================================================
                coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const              0.003548         0.046834            0.076           0.940
L1.Positive       -0.003738         0.003968           -0.942           0.346
L1.Negative        0.005768         0.003616            1.595           0.111
L1.Neutral        -0.001406         0.001540           -0.913           0.361
L1.GME            -0.002385         0.025855           -0.092           0.927
L1.SP500           0.000480         0.006369            0.075           0.940
L1.CPIAUCSL        0.000471         0.799384            0.001           1.000
L1.UNRATE          0.380452        18.399372            0.021           0.984
================================================================================
```

Fig. 4.3.3

```
Best model found at lag 1 with AIC: -3.913821474663611
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
```

Fig. 4.3.4

```
Correlation matrix of residuals
           Positive  Negative   Neutral       GME     SP500  CPIAUCSL    UNRATE
Positive    1.000000  0.993756  0.990527  0.012423 -0.004661  0.000578  0.000411
Negative    0.993756  1.000000  0.990170  0.006176 -0.005521  0.001972 -0.001974
Neutral     0.990527  0.990170  1.000000 -0.000029 -0.014059  0.001093  0.003590
GME         0.012423  0.006176 -0.000029  1.000000 -0.161532  0.049362 -0.084000
SP500      -0.004661 -0.005521 -0.014059 -0.161532  1.000000  0.219419 -0.373143
CPIAUCSL    0.000578  0.001972  0.001093  0.049362  0.219419  1.000000 -0.587532
UNRATE      0.000411 -0.001974  0.003590 -0.084000 -0.373143 -0.587532  1.000000
```

Fig.4.3.5

Performing an optimal lag test on the code with the use of an AIC test, the best model is found with lag 1, on an hourly time frame. This gives an AIC value of -3.9138. In general, a lower AIC is indictive of a better model fit, the optimal lag order, is the value that minimises the AIC value. In this case, the VAR(1) model is the best from a range of model lags 1 to 15. It is important to note that this model prefers a correlation matrix to show no correlation between the residuals of each equation.

The Correlation matrix of the residual in Fig. 4.3.5, shows there to be correlation between the residuals, ideally the residual will show no correlation. Not that this can suggest there is more information to be obtained, or further variable transformations are needed. It can also be a sign to consider alternative models for the interpretations.

4.3.2 Testing for Normality, Heteroskedasticity, and Autocorrelation.

```
Durbin-Watson stat for equation 1: 1.9997404798486818
Durbin-Watson stat for equation 2: 1.9988450592964546
Durbin-Watson stat for equation 3: 2.000332070120758
Durbin-Watson stat for equation 4: 2.000112680816503
Durbin-Watson stat for equation 5: 2.000064358986381
Durbin-Watson stat for equation 6: 2.0000035686997983
Durbin-Watson stat for equation 7: 2.000003568699799
White test: LM-stat: 17.630823321329363, LM p-value: 0.6722318526470152
Ljung-Box test:     lb_stat  lb_pvalue
1  0.000004   0.998398
Jarque-Bera test: JB-stat: 5842073.063230405, JB p-value: 0.0
```

Fig. 4.3.6

Durbin-Watson Statistics: These Value measure the autocorrelation in the residuals of the model. Closer to 2 indicates no autocorrelation, while significantly lower or higher than 2 show positive or negative correlation, in this case the Durbin-Watson statistics are significantly close to 2, therefore the conclusion of no significant autocorrelation in the residuals can be accepted (Durbin, et al., 1951).

White test is checking for heteroskedasticity within the model, the test holds a null hypothesis that is the errors are homoscedastic. The p-value in this output is 0.9722, which is much greater than the significance level of 0.05, meaning we are not able to reject the null hypothesis, and there is no evidence of heteroskedasticity (Black, et al., 2009).

The Ljung-Box test is the check for autocorrelations, this holds the null hypothesis that there is no autocorrelation. The P-value displayed in fig. 4.3.6, 0.9984, is much greater than 0.05, hence the null hypothesis cannot be rejected which is consistent with the Durbin-Watson test, there is no significant autocorrelation within the residuals.

Checking for normality of the residuals, the Jarque-Bera test. Null hypothesis for the JB test is that the residuals are normally distributed. The p-vale is 0.0 < 0.05, therefore we can reject the null hypothesis, this suggests the residuals in this model are not normally distributed, which might be a concern for the VAR model's assumptions (Korstanje, J. 2021) (Jarque, C.M., Bera, A.K., 1987).

4.3.3 Forecasting

```
print(forecast_values)
```

```
[[ 2.26093381e+00  1.35706033e+00  6.52736573e+00  2.85546660e-03
   3.80178175e-02  1.27038626e-03 -9.95600515e-05]
 [ 2.51030997e+00  1.69911715e+00  7.03557147e+00  2.98782411e-03
   3.85803151e-02  1.26620015e-03 -9.92319865e-05]
 [ 2.74884875e+00  2.02689032e+00  7.52556613e+00  3.14099033e-03
   3.91200455e-02  1.26323381e-03 -9.89995147e-05]
 [ 2.97710778e+00  2.34104913e+00  7.99815069e+00  3.28344449e-03
   3.96621777e-02  1.26035649e-03 -9.87740197e-05]
 [ 3.19561652e+00  2.64223107e+00  8.45408965e+00  3.41578396e-03
   4.02058165e-02  1.25756584e-03 -9.85553168e-05]]
```

Fig. 4.3.7

The Forecasting here represents the lag model's forecasted values for a given number of steps, if we take the first row:

Variable	**Forecasted 1st Step Value**
Positive Sentiment	2.26093381
Negative Sentiment	1.35706033
Neutral Sentiment	6.52736573
GME stock close	0.00285547
S&P 500 close	0.03801782
CPIAUCSL	0.00127039
Unemployment Rate (UNRATE)	-0.0000996

Fig. 4.3.8

These forecasted values, Fig. 4.3.8, are based on differenced data, hence, to get actual forecasted value the data requires the researcher to reverse the difference and adding the last know value from the original dataset value. The next steps are optional, the researcher finds it necessary to evaluate the accuracy through Mean Squared Error (MSE) and or Mean Absolute Error (MAE).

	Positive	Negative	Neutral	GME	SP500	CPIAUCSL	UNRATE
0	0.107017	1.172578	1.357053	0.017081	-0.008513	0.001300	-0.000102
1	0.211989	1.340383	1.705949	0.016933	-0.007152	0.001296	-0.000102
2	0.315149	1.503504	2.047131	0.016879	-0.006086	0.001293	-0.000101
3	0.416609	1.662156	2.380893	0.016820	-0.005028	0.001291	-0.000101
4	0.516471	1.816539	2.707519	0.016755	-0.003980	0.001288	-0.000101

Fig. 4.3.9

Forecasting the next 5 steps with the hourly difference data yields the result shown in Fig. 4.3.9. This can be interpreted as the period are in hours and with 5 steps, the values are the forecasted change in the next 5 hours. Note this is not the actual future values for the data, this is tested data with 5 prior steps from the data to test from.

From the table we can infer the Positive sentiment is expected to gradually increase over the next 5 hours, Negative sentiment is also expected to increase, however at a much higher rate comparatively to the Positive sentiment. GME stock returns are expected to stay relatively stable, with only minor fluctuations.

For the financial baseline data, S&P 500 returns are expected to increase gradually, CPI is expected to have small increase and the Unrate is expected to have negligible changes.

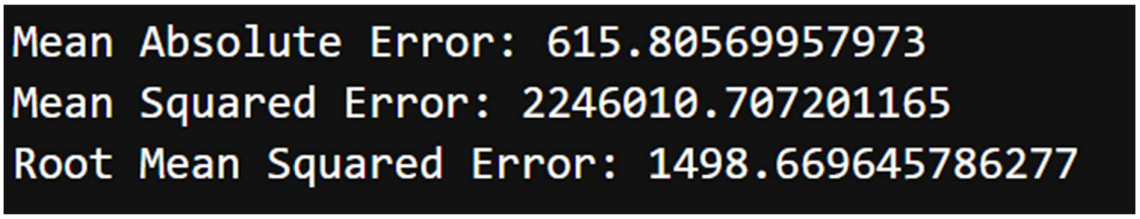


Fig. 4.3.10

The MAE is the average of the total difference between the research forecasted data and the actual values, this measures the average error in each prediction in the same unit as the data. In this case the MAE is 615.81 meaning the predictions are off by 1.2998

MSE, similar to MSE however this gives more weight to the larger errors, so this is sensitive to outliers in the data than MAE. In this model 2.24 million units

RMSE, this is the square root of the MSE and it has the same unit values as MSE, making this error metric more interpretable than MSE. In this case the RMSE is 1,498.67 unit.

These errors are alarmingly high for the used data and show a significant issue within the coding, these are not insignificant errors, this has been some clear malfunctions in the data from the transforming processes during the preparing of the code, or through the splitting of the code for testing.

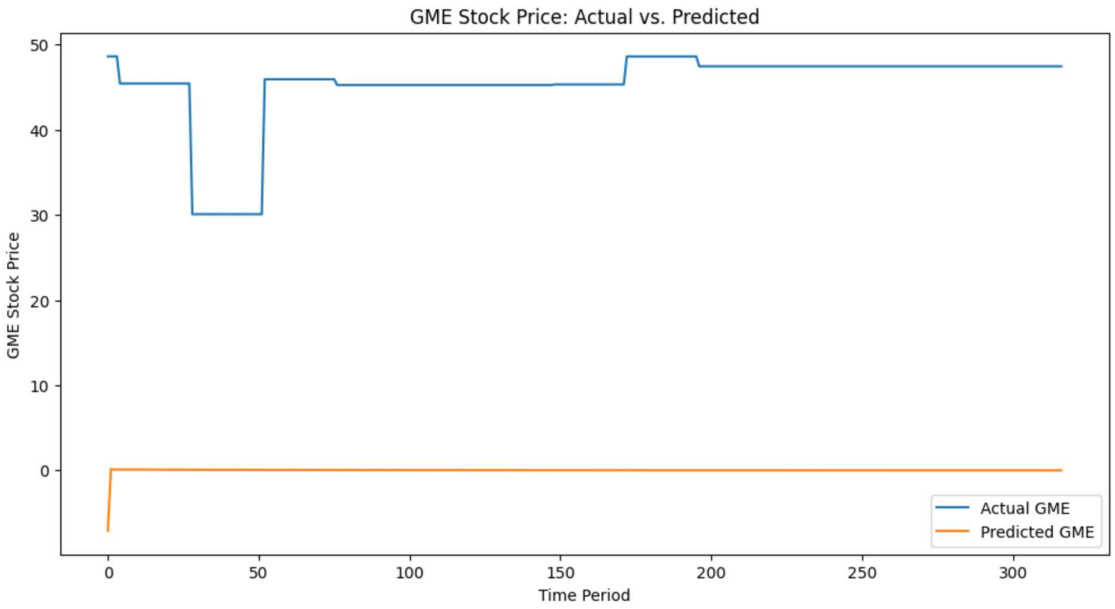


Fig. 4.3.11

This shows the predicted value of the GME stock to almost show no change of improvement over a given time period, comparatively in blue the graph shows the actual price of the GME stock changing actively.

4.4 Summary of Key Findings

The VADER sentiment data shows clear majority in the Positive sentiments being posted in the r/WallStreetBets subreddit. The data also shows to be best optimised when modelling the sentiment released an hour before the stock price is taken, this shows the posts are effective in a shorter time frame, rather than the higher time scales of a daily or weekly movement. The model also finds significant errors in the valuations of the predictive model, furthermore the model test finds a Root Mean Square Error of 1,498.67, this shows the models poor accuracy in predicting the future of the GME stock and other factors are affecting the price in larger ways.

Chapter 5 – Discussion

5.1 Introduction

The discussion chapter opens the question and depth of questioning of the previous chapter's ability to answer the research topic question and objectives. Furthermore, the aims are also to access the certain strengths and weaknesses of the approach the research study took. This chapter will also aim to test the findings to relate to the academic literature.

5.2 Findings in Context to Research Question and Literature

The VADER sentiment data aims to answer the research question, with qualitative, insight into the impact of the positive posts, being the majority of the sentiment analysis posts being posted in the subreddit during January to April 2021 (see Fig. 4.2.4). Visualising the VADER sentiment reveals a trend in the early spike of the GME stock price data (see Fig. 4.2.3) and the sentiment mentions gathered from Reddit, in connection to the literature this is on par with their findings, (Gendron, Y. et al., 2022) (Song, J, 2021), the VADER sentiment shows there to be potential evidence of correlation between the Reddit Sentiment posts and the change in the GameStop stock, this is shown to be a strength of the research approach. The VAR model finds the GameStop stock price, to be inconclusive with the VADER sentiment model, the financial variables are also seeing minimal changes during the shortened time-period. This is in contrary findings to other literature (Gendron, Y. et al., 2022) (Song, J, 2021). The literature suggests the post data collected from the r/WallStreetBets subreddit to be a high factor for influencing the price change in the stock within a larger period.

However, this could be attributed as a weakness and limitation of, the data gathered for this research took place during the short squeeze and did not include the data posts before January 28th, 2021, the data was also tested within the hourly period. An optimal lag test was conducted (Korstanje, J. 2021) and found the best lag for the VAR model to be within one hour, this does show potential correlation between the Sentiment data to be most accurate at testing within the shorter period rather than the range tested, this does link with the literature (Diangson, B. Jung, N., 2021). This could be a result in the data showing a descent from the most successful period of the stock price climb (Yahoo, 2022). This was due to limitations within the researcher's data gathering capacities and scope of the research, however this does allude to potential further research in testing the data before January 2021, to access how sentiment influence gathered before the height of the event may provide changes in the pricing of the stock.

In a broader research question of how the new media influences the financial markets, the notable Robinhood trading platform for retail investors (Fedyk, V., 2022). This is significant to the research findings as the most popular trading platform at the time offering the options for buying GameStop stock, was Robinhood, they forcibly removed the choice for retail traders to open new positions within the GameStop stock (Fedyk, V., 2022), to halt the momentum and losses of the hedge funds shorting the GME stock (Kabir, U., 2021) This in the broader context of the research shows that the new media's effect on the financial markets require further comparisons and data to infer firm causations. The VAR values of the models are stretching into poor modelling and could be a weakness of the research as the data may be

lacking sufficient information or the comparisons for this research are being displayed in the incorrect model. The literature (Song, J., 2021), (Fedyk, V., 2022), (Gendron, Y. et al., 2022), concludes the opposite of the VAR findings within the chapter 4, Song, J., (2021) concludes a significant link between the discussions on the r/WallStreetBets subreddit and the fluctuation with the daily pricing of the stocks gathered. This seems to be the overall agreed connection between the mentioned stock sentiment posts, within Reddit, and the actual daily price change occurring to the price of the stocks.

The final forecasting could have been corrupted from the actual predictions, based on the assumptions being broken for the VAR model, this can be tested for (Korstanje, J. 2021). This was shown in the Jarque-Bera test which holds the null hypothesis for the residuals to be normally distribute (Jarque, C.M., Bera, A.K., 1987), where the assumption for the VAR model is that the data is normally distributed (Korstanje, J. 2021). Jarque-Bera test for this research resulted in a p-value 0.00 which is less the significance level of 0.05, therefore the null hypothesis can be rejected, breaking a key assumption of the VAR model (Korstanje, J. 2021). This coupled with the MAE and MSE results this can highlight if the model is performing well with predicting the given data.

Chapter 4 also revealed the data of the VAR model forecasting Mean Absolute Error (MAE) and Mean Square Error (MSE). The MAE and MSE are measures of the model's ability to perform the predictions, the higher the score the worse the ability of predicting is. The research data used gave MAE and MSE value that are extremely high, (Fig. 4.3.10), for this dataset type, this could be due to poor pre-processing or further missing data, it could also mean the data requires further transformation or that the transformations have resulted in poor data being used (Korstanje, J. 2021). These error values for this data can show how well the model is performing when forecasting the future values, further visual inspection can show the issues within the data sets.

5.3 Integration of Mixed Methods

The combination of the qualitative VADER sentiment and the quantitative financial data variable have been relevant in other literature research and are a strength of the research approach (Gendron, Y. et al., 2022) (Song, J, 2021). The combination of the VADER model and the VAR model give the research question, of how new media effects the financial markets, a depth of insight that would be lost without the VADER sentiment model. The qualitative aspect of the posts on reddit give a way to understand the impact that online forums can generate within in short time frames to affect the real quality of the market environments. Not in a small way, price shows significant spike during the same period as the posts are at peak events, this is a depth of insight that financial data alone would be unable to give. The VAR model was used to forecast the financial variable and sentiment data, the model finds implications through the final steps, showing a poor relationship between the variables, furthermore the models implicated through breaking key assumptions, this leads the researcher to compare the results and findings against existing literature to conclude a different result.

Chapter 6 – Recommendations and Conclusions words

6.1 Limitations and Future Research

In the case study of the GameStop Short Squeeze this research has given a depth of insight into the potential impact that new media can have on the financial markets, especially within a shortened time frame. This research explores the historical posts of Reddit, with the VADER sentiment model findings, during the event, majority of posts were positive and could show momentum in the buying positions of the GME stock. This study is limited through the lack of inclusive data for the Robinhood halting the opening of positions which effected the data after February. Future research could seek to include this and adjust the data collection to be aware of this. Limitations are mentioned throughout the chapters. The scope of the research did not allow for a larger data size to be captured, this coupled with the limitations of the hardware and time constraints the researcher faced, with limited python coding knowledge, created a significant gap in the coding performance, future research could use this as a guide for the logical steps and potential highlighted issues when conducting their own research. VADER sentiment finds limitations with being unable to make use of videos and pictures which is commonplace on reddit and if new media continues to provide financial insight, this may be a critical step to overcome.

6.2 Conclusion

To conclude, the research question of whether new media effected financial markets was explore and interpreted to great length within the scope and limitations of the research and researcher. New media, specifically the subreddit r/WallStreetBets, has shown to be highly active within the date event of the GameStop Short squeeze. It can be concluded that most posts within the subreddit during the time were positive towards the GameStop (GME) stock. New media has shown to bring influence on the financial markets, this could also be linked with COVID-19 as the influx of people staying at home caused a rise within the retail traders. However, after the peak in January the mentions and pricing of GME descended significantly. The VAR model showed there to be potential signs of the financial variables having effects however, through testing the VAR model for the assumptions after transforming, the data showed to be inconclusive, requiring further transformations outside of the scope of this research. The issues with the VAR model could likely stem from several possible factors, one of which is the presence of coding issues, this may have led to inaccurate or incomplete data being portrayed in the analysis and findings a lack of correlation between the VAR analysis and the literature. This model may also be inappropriate capture the true relationship between the variables.

These issues of unexpected findings in the analysis bring implications for this study and the broader field. For example, this research highlights a need for a deeper investigation into the methods and models that are used for conducting a thorough study on financial markets and new media. The efficient market hypothesis could also prove to be more complex and accomplished than initially presumed.

6.3 Final Thoughts

This study contributes to the new media's role in financial markets and the EMH. The analysis gives insight to the trends and complex nature between financial markets and new media, emphasising the need for future research to expand on the questions within this study and a deeper refined methodology.

This study has brought light to interesting questions with new media, financial markets, and efficient market hypothesis exposing trends and uncovering patterns through sentiment analysis. An unexpected outcome from the VAR shows the need for an expanded and deeper delve into the methodology for future research. Research that can address the limitations and build upon the work in this research can continue the advancement of knowledge of the relationship held between financial markets and new media.