CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Gamified Stock Markets, Sentiment and Volatility: Evidence from the GameStop frenzy

Master's thesis

Author: Bc. Thai Nhat Phi Tran Nguyen

Study program: Economics and Finance

Supervisor: prof. PhDr. Ladislav Krištoufek, Ph.D.

Year of defense: 2022

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, 1st August 2022

Thai Nhat Phi Tran Nguyen

Abstract

In this thesis, we study the impact of individual retail investors on the financial markets. We follow the GameStop retail trading frenzy from the beginning of 2021 and the retail investors aggregated on Reddit's *r/wallstreetbets*. The tools employed include natural language processing, wavelet analysis and vector error correction models. The results propose that the retail investor sentiment is highly susceptible to high volatility, extreme returns and frequent news coverage. Social media is shown to exacerbate these behavioural tendencies. We find evidence that retail investor sentiment is able to predict short-term returns for stocks specifically targeted by retail investors. The findings are, however, dependent on the investment horizon. Over long horizons, we find evidence for the reversal of the relationship. Lastly, while the effect of news and social media is similar in the long run, we show that Reddit sentiment, as opposed to news sentiment, is a significant predictor of retail targeted stocks in the near term.

JEL Classification	C55 C58, G12, G14, G41
Keywords	Sentiment, Social media, GameStop, Reddit, Natural language processing, Wavelet analysis
Title	Gamified Stock Markets, Sentiment and Volatility: Evidence from the GameStop frenzy

Abstrakt

Hlavním předmětem této práce je studie vlivu jednotlivých investorů na finanční trhy. Konkrétně následujeme ságu kolem akcií společnosti GameStop ze začátku roku 2021 a retailové investory, kteří se shromáždili na fóru *r/wallstreetbets* na sociální síti Reddit. Mezi použité nástroje patří zpracování přirozeného jazyka, vlnková analýza a vektorový model korekce chyb. Výsledky naznačují, že zejména vysoká volatilita, extrémní výkyvy v cenách či časté pokrytí zprávami lákají drobné investory. Dále se ukazuje, že sociální média tyto tendence chování ještě zesilují. Nacházíme zde důkazy, jež naznačují, že sentiment drobných investorů je schopen predikovat krátkodobé výnosy akcií, které drobní investoři specificky zaměřují. V dlouhých horizontech nicméně dochází k obrácení vztahu mezi sentimentem a výnosy. Na závěr, zatímco v dlouhodobém horizontu je efekt sentimentu zpráv a sociálních medií shodný, tak sentiment Redditu, na rozdíl od sentimentu zpráv, je v krátkodobém horizontu významným faktorem ovlivňujícím akcie zaměřené drobnými investory.

Klasifikace JEL	C55 C58, G12, G14, G41
Klíčová slova	sentiment, sociální média, GameStop, Reddit, zpracování přirozeného jazyka, vlnková analýza
Název práce	Akciové trhy jako hra: Nálada a volatilita během GameStop horečky

Acknowledgments

I wish to express my utmost gratitude to prof. PhDr. Ladislav Krištoufek, Ph.D.for providing me with the opportunity of writing this project as well as the provided valuable council and appreciated encouragement. Further, I am grateful for the valuable and insightful comments and advice by PhDr. Jiří Schwarz Ph.D. and prof. PhDr. Tomáš Havránek Ph.D. Lastly, a wholehearted appreciation also goes out to my family and friends for their continuous support and for being a source of inspiration.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Tran Nguyen, Thai Nhat Phi: *Gamified Stock Markets, Sentiment and Volatility: Evidence from the GameStop frenzy.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2022, pages 80. Advisor: prof. PhDr. Ladislav Krištoufek, Ph.D.

Contents

List of Tables **3**

List of Figures **4**

Acronyms **5**

Introduction **11**

1 Literature Review **16**

1.1 Noise Trader Theory 16

1.2 Individual Investor Sentiment, News & Social Media 17

1.3 News & Social Media Sentiment 20

1.4 GameStop Related Literature 21

2 Methodology **23**

2.1 Text processing 23

2.2 Sentiment extraction 23

2.3 Wavelet analysis 24

2.3.1 Applied literature 24

2.3.2 The continuous wavelet transform 26

2.3.3 Wavelet coherence 26

2.3.4 Phase 27

2.4 Vector error correction model 27

3 Data **31**

3.1 Textual and sentiment data 31

3.2 Financial data 35

4 Results **39**

4.1 Text analysis 39

4.1.1 Sentiment analysis 40

4.2 Wavelet Coherence 42

4.3 Comovements 46

4.4 Vector error correction model 49

4.4.1 News Sentiment 53

5 Discussion **56**

6 Conclusion **60**

Conclusion **60**

Bibliography **71**

Appendix **72**

List of Tables

1 Sample of Reddit posts from January 2021 34
2 Example of sentiment scores 34
3 Summary: Sentiment 36
4 Summary: Returns 37
5 Summary: Volatility 38
6 Stationarity tests 50
7 Johansen cointegration tests: Sent 51
8 VECM results: Sent 51
9 Error correction terms: Sent 52
10 Granger causality: Sent 53
11 Johansen cointegration test: news 54
12 VECM results: news 54
13 Granger causality: news 55
1 A.1 - Granger causality: daily Sent 74

List of Figures

1 GME Relevancy on WSB 32
2 Difference between on- and off-trading hours sentiment 33
3 Comparison of sentiment and market activity 39
4 Word and Bigrams frequency 40
5 Ticker mentions 41
6 Sentiment and volatility heatmaps 42
7 Wavelet analysis: Reddit Sentiment 43
8 Wavelet analysis: News and Twitter sentiment 45
9 Wavelet analysis: Comovements 47
10 Wavelet analysis: Broad market and short interest index 48
1 Summary of stock returns 72
2 Sentiment distribution 73
3 Sentiment comparison 73
4 Wavelet analysis: Volatility spillovers 75

Acronyms

GME GameStop

AMC American Multi-Cinemas

BB BlackBerry

CLOV Clover Health Investments Corp

NOK Nokia

RKT Rocket Companies Inc

Sent Reddit retail investor sentiments

WSB wallstreetbets

VECM Vector error correction model

DSSW De Long, Shleifer, Summers and Waldmann

AIC Akaike information criterion

BIC Bayesian information criterion

HQ Hannan-Quinn information criterion

FPE Final Prediction Error

VAR Vector autoregression

VADER Valence Aware Dictionary and sEntiment Reasoner

NLTK Natural Language Toolkit

CLDR Common Locale Data Repository

API Application Programming Interface

IPO Initial Public Offering

NYSE New York Stock Exchange

CEF Closed-end Funds

SAD Seasonal affective disorder

EMH Efficient market hypothesis

Master's Thesis Proposal

Author	Bc. Thai Nhat Phi Tran Nguyen
Supervisor	prof. PhDr. Ladislav Krištoufek, Ph.D.
Proposed topic	Gamified Stock Markets, Sentiment and Volatility: Evidence from the GameStop frenzy

Motivation The events surrounding GameStop at the beginning of 2021 gained a lot of traction when the price shot up more than 2000% in January. Six months later, the price runs above $200, about 12 times its closing price on the first day of 2021. The surge of GameStop and other stocks is unique in the aspect that it originates from a trading forum (called a subreddit) on the social media site Reddit. A large amount of small retail investors on the mentioned platform cooperated to thwart substantial short positions in GME (and others). Effectively, the movement constitutes an attempt to crowdsource a short squeeze. The collusion and involvement of many individual investors offer unique optics for the analysis of investor sentiment and its impact on prices and volatility. The motivation of the thesis is split three ways, revealed below.

Retail investor & Signalling quality of information The rise of retail investing paired with a copious amount of rapidly disseminated information raise a question about its signalling quality. Processing, respectively interpreting, the vast ocean of information may possess a quality signal, respectively insight, about the financial markets. The focus being on retail investors, this insight may shed a light on the degree of sophistication of the seemingly "uninformed" retail investors. Specifically, how does sentiment alter investment decisions. We propose to examine the above through the optics of the retail trading frenzy surrounding GameStop. The role of noise traders with respect to returns, volatility and sentiment has been keenly studied, see De Long et al. (1990) or Shleifer and Summers (1990).

Sentiment & Medium of information

The effect of sentiment-driven behaviour is long documented see Brown (1999), Barber and Odean (2008), or Baker and Wurgler (2006). More recently, the distinction between the effect of news media and social media has been scrutinised (Jiao et al., 2020). Given the role of social media, its distinct effect on retail investor sentiment is of interest. Particularly, with comparison to overall and news sentiment surrounding financial markets.

Market manipulation Unsurprisingly, the notion of manipulative trading is predominantly associated with institutional traders, respectively "large traders" possessing the ability to move prices. Anomalously, the GameStop saga possibly documents a rare case of cornering the market by the individual retail investors who orchestrated the movement, largely through social media. Contrastingly to the usual perception of market manipulation as a scrap between Wall Street names, the GameStop saga offers a unique affair driven by Main Street against Wall Street

sentiment. However rare, this is not the first attempt of market cornering motivated by similar sentiment. The parallel dates back to the first supermarket chain Piggly Wiggly in the 1920s, the formal overview of which is given by Allen et al. (2006). Alternatively, for a more interesting and informal take refer to this twitter feed. Ultimately, the collusion and mobilisation of a large amount of individual retail investors offer distinctive optics on the subject of possible market manipulation and predatory pricing.

Market contagion While GameStop remains at the centre of the retail investor mania, its effect was not contained in the GME stock. Some spillover effects have been suggested, for stocks such as AMC or BlackBerry (BB), see Tengolov et al. (2021). Therefore, in addition to GameStop, we bid to examine other targeted stocks, particularly those mostly mentioned in the WSB subreddit. Considering only a small subsample of stocks, we do not expect to examine the financial market contagion of the GameStop mania, something already covered by Aharon et al. (2021).

Hypotheses

Hypothesis #1: The GameStop saga was a result of predatory pricing from retail investors.

Hypothesis #2: Individual investor sentiment has had a significant impact on stock's volatility and volume.

Auxiliary hypothesis#2.1: Sentiment from social media is distinctively different from news sentiment.

Hypothesis #3: The effects of retail trading frenzy were not exclusively contained in GameStop.

Methodology

Text mining and sentiment analysis To analyse the sentiment and motivation behind the GameStop mania, we need not go further than Reddit investing forums, where it originated. Using Reddit's API we access the data on specific posts, daily discussions, and relevant mega threads. For textual analysis and sentiment extraction, Python's NLTK library is utilised. Additionally, for comparison, exogenous measures of sentiment are used, such as News Sentiment datasets from WorldData.AI, or sentiment surveys from AAII.

Stock and options data Historical stock quotes can be obtained from Yahoo Finance. Using automated scripts, intraday data from Yahoo Finance is downloaded, however, only from May onwards. Similarly, data on GME option chains are downloaded each Friday before the open. Short sale volume data is available directly from FINRA's bi-monthly reported Short Sale Transaction Files, or from Quandl's streamlined dataset for individual tickers. Furthermore, specific data on loan fees and stock availability is available from Interactive Brokers API.

Wavelet Analysis The utilisation of the wavelet framework is motivated by its ability to capture information from both time and frequency domains. We conjecture that the wavelet analysis will be better suited for the complex dynamics of the relationship between sentiment, volatility, prices and volume. The framework allows not only analysis of the impact of sentiment on relevant stocks but also their comovements. The latter will offer insight into the relative market contagion of the retail investor frenzy. Regarding its adequacy and employment refer to Torrence and Compo (1998). The use of wavelets in this context is not novel, see Long et al. (2021) or Umar et al. (2021) who utilise the wavelet analysis to assess the GameStop saga.

Time-Series Analysis In addition to the wavelet framework for modelling the comovements, we can resolve to standard time-series tools. Particularly, employing the models from the multivariate GARCH family, i.e. DCC-GARCH or the more complex cousin BEKK-GARCH, which allows for time-varying volatility, introduced by Engle (2000) and Engle & Kroner (1995), respectively.

In summary, the first hypothesis is tested using semantic analysis via extracting textual information and sentiment from social media posts. Concerning the auxiliary hypothesis, comparisons with news sentiment will be drawn. To examine H2 both wavelet and traditional time-series analysis can be used. Specifically, extracting the significant relationships, respectively comovements, between sentiment, prices and volume are of interest. Lastly, the analysis of comovements offers an avenue to explore contagion among the relevant stocks, i.e. those most mentioned with GME.

Expected Contribution Focusing on the GameStop saga from the beginning of the year, the thesis aims to contribute to the growing body of literature examining the specifics of this event through various optics. Following the structure from the first section, the expected contribution is three-fold. First, the thesis intends to contribute to the discussion regarding the impact of retail investors, specifically the role of individual investor sentiment. Further, the topic also lies within the literature on market manipulation or predatory pricing. Contrary to the norm, the thesis attempts to illustrate a rare case of market manipulation orchestrated by individual investors and essentially crowdsourced short squeeze. Lastly, using the novel wavelet coherence analysis the thesis will examine the contagion of the retail investor frenzy among related stocks.

Limitations Now, let us briefly discuss what the study is NOT about. While the GameStop frenzy and sentiment surrounding is the focal point of analysis, we do not intend to discuss the socioeconomic and strictly behavioural optics of this event. Following that, the involvement and the roles of hedge funds, market makers and other institutions are adjacent to the analysis and not of main interest. Lastly, the thesis' aim is not to advocate/argue policy implications of the GameStop saga with respect to regulation and market efficiency.

Outline

1. Introduction

2. Literature Review
    - Noise trader theory
    - Sentiment and volatility
    - Market manipulation
    - Comovements using wavelet analysis
3. Methodlogy
    - Textual & sentiment analysis
    - Wavelet & spectrum analysis
    - Multivariate GARCH
    - Hypotheses
4. Data
    - Social media
    - News & overall sentiment
    - Stock and options
    - Short sale volume
5. Sentiment
    - Textual analysis
    - Sentiment analysis
    - Interactions between sentiment and stock price movements/volatility
6. Time-Frequency Analysis
    - Spectrum analysis
    - Wavelet coherence (Comovements)
7. Multivariate GARCH (pending)
    - BEKK/DCC-GARCH
8. Results & Discussion
    - Sentiment Analysis and impact of investor sentiment on volatility and volume
    - Comovements and contagion
    - Causality
9. Conclusion
    - Implications
    - Limitations

Core bibliography

Aharon, D. Y., Kizys, R., Umar, Z., & Zaremba, A. (2021). Did David Win a Battle or the War Against Goliath? Dynamic Return and Volatility Connectedness between the GameStop Stock and the High Short Interest Indices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3788155

Brown, G. (1999). Volatility, Sentiment, and Noise Traders. Financial Analysts Journal, 55(2), 82-90. Retrieved June 13, 2021, from http://www.jstor.org/stable/4480157

Chohan, U. W. (2021). Counter-Hegemonic Finance: The Gamestop Short Squeeze. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3775127

De Long, J., Shleifer, A., Summers, L., & Waldmann, R. (1990). Noise Trader Risk in Financial Markets. Journal of Political Economy, 98(4), 703-738. Retrieved June 13, 2021, from http://www.jstor.org/stable/2937765

Engle, R. F. (2000). Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.236998

Engle, R., & Kroner, K. (1995). Multivariate Simultaneous Generalized Arch. Econometric Theory, 11(1), 122-150. Retrieved June 16, 2021, from http://www.jstor.org/stable/3532933

Fusari, N., Jarrow, R., & Lamichhane, S. (2020). Testing for Asset Price Bubbles using Options Data. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3670999

Hasso, T., Müller, D., Pelster, M., & Warkulat, S. (2021). Who participated in the GameStop frenzy? Evidence from brokerage accounts. Finance Research Letters, 102140. https://doi.org/10.1016/j.frl.2021.102140

Jiao, P., Veiga, A., & Walther, A. (2020). Social media, news media and the stock market. Journal of Economic Behavior & Organization, 176, 63–90. https://doi.org/10.1016/j.jebo.2020.03.002

Long, C., Lucey, B. M., & Yarovaya, L. (2021). I Just Like the Stock versus Fear and Loathing on Main Street : The Role of Reddit Sentiment in the GameStop Short Squeeze. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3822315

Pedersen, L. H. (2021). Game On: Social Networks and Markets. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3794616

Shleifer, A., & Summers, L. (1990). The Noise Trader Approach to Finance. The Journal of Economic Perspectives, 4(2), 19-33. Retrieved June 13, 2021, from http://www.jstor.org/stable/1942888

Tengulov, A., Allen, F., Nowak, E., & Pirovano, M. (2021). Squeezing Shorts Through Social News Platforms. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3823151

Torrence, C., & Compo, G. (1998). A Practical Guide to Wavelet Analysis. Bulletin of the American Meteorological Society, 79, 61-78.

Van Wesep, E. D., & Waters, B. (2021). The Sky's the Limit: Asset prices can be indeterminate when margin traders are all in. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3785637

Umar, Z., Gubareva, M., Yousaf, I., & Ali, S. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. Journal of Behavioral and Experimental Finance, 30, 100501. https://doi.org/10.1016/j.jbef.2021.100501

Introduction

The times when financial markets were playgrounds reserved for and dominated only by the institutional investors are long gone, as besides the sophisticated investment industry, retail investors, or often pejoratively so-called *dumb-money*, enter the equation. The increased access to financial markets and overall investing paired with near-zero commissions on execution have led to a meteoric rise in retail investing. According to Bloomberg Intelligence compiled data, retail investors accounted for nearly a quarter of traded volume in US equities in 2021, representing a marked increase compared to the stable trend in the previous decade [1]. Therefore, the impact of retail investors is no longer merely a side thought which can be quickly glossed over, even by the institutional investors, as evident from the comments of industry professionals and research presented by Financial Times or Deloitte (Martin and Wigglesworth, 2021 [2]; Deloitte, 2021 [3]). Even though the marked increase in retail investing is the captivating narrative of recent times, the impact of individual investors has long been studied, as shown by many academic works from the past, see Odean (1999), Barber and Odean (2002), Kumar and Lee (2006) or Barber et al. (2009) [4, 5, 6, 7], among many others, which will be presented later in the relevant section. Interestingly, it needs to be said, their impact was found highly significant. Presumably, with many market frictions greatly reduced or completely out of the way, the impact of individual investors might be even more pronounced given these conditions.

The advances in technology not only improve access to financial markets but also, perhaps, more importantly, access to information. Nowadays, there is a copious amount of information available, rapidly disseminated within several clicks. The quality of such information is, however, largely up in the air, which raises the question regarding its signalling quality. The processing and interpretation of this vast ocean of knowledge, therefore, presents an important avenue as it, presumably, slightly levels the playing field between the seemingly uninformed, retail investors and the sophisticated, institutional investors. From this point of view, the rise of social media plays a vital role, provided that it represents a platform, where users can not only share relevant information to wider investor masses, but also discuss contemporaneous market events, and strategies or even seek investment advice. We believe that from the interactions amidst the masses of individual investors arises an important and quality signal concerning the retail investor sentiment and their impact on financial markets. Admittedly, this is not an unconventional notion, as shown by early work of Antweiler and Frank (2004) [8] or more recent by Chen et al. (2014), Jiao et al. (2020) or Farrell et al. (2021) [9, 10, 11]. Furthermore, exploring the individual investor sentiment may shed a light on the degree of sophist-

ication of the *dumb-money*. Given the focus on retail investors and highlighted role of social media, we dedicate our thesis to the examination of the signalling quality of retail investors' sentiment, with a particular focus on social media, through the optics of the retail trading frenzy closely associated with GameStop. Specifically, our study is primarily framed within the *noise trader theory* of Black (1986) [12] and further extended by Shleifer and Summers (1990) [13], which essentially postulates that uninformed traders, or so-called noise traders, driven by sentiment, can affect asset prices and effectively lead them to deviate from the fundamentals.

The events surrounding GameStop at the beginning of 2021, at the core of which were Reddit retail investors, gained a lot of traction and attention from investor masses, including institutional investors, the general public and even regulators when the price of the stock skyrocketed and rose nearly twentyfold within the first month of 2021. Twelve months, more than a quarter of a million direct mentions on Reddit, and a couple of congressional hearings later, the price still hovered around ten times its closing price on the first day of 2021. There are two major identifying features of this GameStop saga by which our particular focus on this period is motivated. First, the undoubtedly substantial role of individual investors aggregated on a trading forum, called subreddit, on the social media platform Reddit. In particular, the unusual cooperation and large formation of retail investors whose activity closely accompanied the exorbitant surge in the GameStop (GME) stock as well as other most mentioned stocks alongside it. Second, the initial motivation behind the collusion and involvement of a large number of retail investors which was in reaction to extensive short positions in GME they aimed to thwart. More than the intentions and motivation of the retail investors in this saga, we are interested in the limited arbitrage aspect, given the extensive short-interest and extreme price volatility resulting in more expensive arbitrage. Essentially, the retail trading frenzy constitutes an attempt to crowdsource a short squeeze driven by investors aggregated on a social media platform. The involvement of a large group of retail investors, in addition to social media playing a cardinal role in the saga, offers highly unique optics for the analysis of investor sentiment and its impact on prices and volatility.

As alluded to, we focus on Reddit retail investors, in particular the *r/wallstreetbets* subreddit users, and obtain data on the relevant submissions on the forum throughout the year 2021. Analysing more than 450,000 submissions using natural language processing, we extract key information concerning investor sentiment represented as sentiment scores. In addition to these scores, which present the core data for our analysis, we also employ daily news and Twitter sentiment data series downloaded from Bloomberg, over the period November 2020 - November 2021. Alongside GME, we identify other relevant tickers most mentioned by the Reddit investors at the beginning of 2021. For these stocks, we obtain

high-frequency tick-by-tick 30-min intraday stock quotes, in addition to daily quotes data considering the sampling frequency of the Bloomberg sentiment data. Apart from the textual analysis and sentiment extraction from the natural language processing toolkit, our analysis is built on two main methods. First, the bivariate wavelet coherence framework, used to analyse dependencies between the variables in both time and frequency domains. Second, primarily motivated by the results of the former and to further complement them, we employ a traditional time-series tool in the form of a vector error correction model, which provides us with apparatus to statistically infer and test causality between the variables. We suspend the description of the data and methods here. The detailed description of the employed tools is due later in the relevant sections.

We begin the analysis by obtaining sentiment from the scraped subreddit posts, which, apart from containing a substantial amount of noise, indicate overwhelmingly positive sentiment throughout the sample period, in contrast to news and, interestingly, even Twitter sentiment. Both of which tend to at least oscillate between positive and negative moods. Textual analysis and sentiment scores suggest that social media, particularly Reddit, exacerbate the already present behavioural tendencies of retail investors, including overconfidence, self-attribution bias or confirmation bias. Moreover, the proclivity of retail investors to trade similarly is strongly reinforced by the discussion on the subreddit, effectively leading to an even stronger and more uniform individual investor impact.

The wavelet analysis shows large and significant areas of dependence between sentiment and returns, albeit using high-frequency data these periods do not last long. However, they coincide with large price volatility spikes in the respective stocks. Thus, showing that the appeal of high volatility, extreme one-day returns and news coverage play a major role in drawing retail investors' attention, as documented by Barber and Odean (2008) [14]. Contrary to Atkins et al. (2018) [15], we find the coherence between sentiment and returns stronger than for volatility, although sentiment and volatility exhibit a significant positive relationship, further substantiating the individual investors' attraction to volatility. We find similar results for news sentiment, but while news sentiment appears to be largely driven by GME in the short-run, Reddit sentiment possesses predictive power for the retail targeted stocks. Over long horizons, we find that the relationship between stock returns and retail investor sentiment turns opposite, suggesting that, while individual investor sentiment is able to predict short-run dynamics for brief periods, the relationship tends to reverse in the long run as price correction occurs.

Lastly, we show that the retail trading frenzy was not contained to GME but also extended to other mentioned stocks, some of which exhibited the same features as GME, consequently drawing the attention of retail investors. Specifically, we detect large and significant comovement between GME and the alongside mentioned stocks that the Reddit

investors likewise targeted, with AMC being the particular standout. Further, AMC is also shown as a significant driver of sentiment and other relevant stocks, in contrast to GME, which is, interestingly, found to largely react in response, as indicated by Granger causality in the short-run. Notably, lower frequency sampling reveals that GME was also a major determinant of sentiment and other affected stocks. Nonetheless, in addition to strong wavelet coherence indicating large comovement, we detect a long-run cointegrating relationship.

We aim to contribute to the growing body of literature which navigates the current unique market accessible to wider investor masses with next to no market frictions and focuses on individual investors. The role of individual investors, respectively their impact, has long been studied, although, we believe that given current conditions, further empowering retail investors and levelling the field in the hassle between so-called *dumb* and *smart* money, their role ought to be examined amidst these factors. Furthermore, as economics and financial markets become increasingly grappled by viral stories, overarching opinions and beliefs, for which R. Shiller [16] coined the term *narrative economics* describing this phenomenon, the outstanding importance of sentiment is highlighted. The core contribution of our thesis is essentially split three ways. First, analysing individual investor sentiment, especially related to the compelling events surrounding GameStop, we describe the behaviour and sentiment of this particular breed of investors who aggregated on Reddit, within the context of the largely documented behavioural tendencies of retail investors and the fact that their behaviour largely converges. Second, our work intends to illustrate the significant impact of sentiment on returns and volatility of stocks, which are targeted or heavily discussed among retail investors. Furthermore, our employed methods allow us to examine how the effect sentiment varies across time and frequency and distinguish between short- and long-run dynamics. Thirdly, we mainly discuss the previous two points within the context of social media sentiment. One of our main interests is to highlight the role of social media, specifically, its role in empowering individual investors, thanks to the discussion, information sharing etc., exacerbating behavioural tendencies and intensifying sentiment. Moreover, we aim to advance the recent literature which emphasises the distinct value of social media within the context of financial markets and carries the general idea that social media sentiment, while different from traditional news sentiment, conveys valuable information. In sum, our thesis aims to discuss the individual investors' biases and their ability to affect financial markets in both short- and long-run, with a particular focus given to the role of social media concerning the augmentation of the former and bolstering of the latter.

The rest of the thesis is organised as follows. Section 1 presents related literature, comprising the theoretical background of our study and relevant empirical findings, in-

cluding some of the recent academic works which likewise covered the GameStop retail trading saga. Next, Section 2 describes the employed empirical apparatus central to our analysis. Section 3 contains a brief discussion regarding the utilised data and a short preliminary analysis of the data. Section 4 contains the cardinal results of our empirical methods. Further discussion concerning the outcomes and the contribution of our work is provided in Section 5. The thesis' principal findings and contribution are subsequently summarised in Section 6.

1 Literature Review

Following our motivation, we categorise the literature survey into three general sections: noise trader theory, investor sentiment, and news and social media sentiment. Lastly, in addition to these strands of literature, we separately discuss numerous recent academic works which, similarly, cover the GameStop retail trading saga, as it gained plenty of attention from investors, outsiders and academia alike.

1.1 Noise Trader Theory

Given the thesis' focus on the individual investor, their sentiment and the overall nature of the GameStop trading frenzy, the following section motivates our departure from the efficient market hypothesis, the overview of which can be found in the seminal work of Fama (1970) [17]. A more recent survey of the theory is found in Yen and Lee (2008) [18]. Related to the emphasis on the role of irrational agents, trading on information and noise, and imperfect, respectively limited, arbitrage, we present the strand of literature within the *noise trader theory*. Thus, effectively allowing the possibility of long-term deviations from the fundamentals.

However, firstly, there have been numerous stylised facts, respectively financial anomalies, backed by extensive empirical evidence irreconcilable with the theory of efficient markets, some of which we discuss here to motivate our departure from the EMH. Including, but not limited to, Shiller's proof of excess volatility [19], Bond and Thaler's (1985) overshooting [20] and day-of-the-week effect, among many others. Further, the *January effect* or *Turn-of-the-year effect*, first documented by Rozeff and Kinney (1976), [21] describes the empirical finding that small stocks generate significantly higher returns than their large counterparts and market indices in January, especially following a decline in December. Roll (1983) and Reinganum (1983) [22, 23] ascribe the effect to the tax-loss-selling, creating price pressures on stocks that have poorly performed throughout the year. After the turn-of-the-year, this pressure is alleviated and returns revert to equilibrium. Their hypothesis is further generalised by Ritter (1988) [24] who emphasises the role of buying and selling behaviour of individual investors, particularly the reinvesting in January following the tax-loss-motivated selling before the end of the year. The "January effect" is one of the many market inefficiencies which document prices driven by factors other than fundamental news. In theory, given the existence of non-constrained arbitrage, it could not exist as temporary trading patterns would be eliminated.

Following Kyle's (1985) [25] definition of random noise trader, Black (1986) [12] establishes the *noise trader theory* representing a distinct departure from the theory of efficient

markets. Black [12] divides traders into information and noise traders, the latter of which trades on noise as if it were information, but ultimately both groups do not know for certain whether they are trading on information or noise and whether or not the information had already been priced in. Moreover, two reasons for the prevalence of noise trading are provided. First, investors plainly like to trade on noise, and second, they believe they are trading on information.

In reaction to the documented empirical shortfalls of the efficient market hypothesis, Shleifer and Summers (1990) [13] argue in favour of pursuing an alternative approach, one based on Black's (1986) [12] model, motivated by two central themes. First, the existence of limited arbitrage, due to its inherent riskiness, which extends beyond the fundamental risk and also includes the probability that the mispricing becomes even more extreme in the future, described as *future resale price risk*. Second, the existence of irrational agents whose demand for risky assets is impelled by their beliefs and sentiment irrespective of fundamentals. De Long, Shleifer, Summers and Waldmann (hereafter DSSW) in a series of works (1989, 1990a, 1990b) [26, 27, 28] further expand on the *future resale price risk* and the unpredictability of investor sentiment, described as the *noise trader risk*. They develop a model in which sophisticated investors and noise traders interact, where the latter trade on incorrect beliefs and pseudo-signals. In addition to being able to explain some of the financial anomalies such as Mehra and Prescott's (1985) [29] equity premium puzzle or undervaluation of closed-end mutual funds, it is shown that noise traders can earn higher expected returns from bearing the noise trader risk they create.

Nonetheless, as noted above, it is not in our interest to directly address the efficiency of the markets or test the efficient market hypothesis. Instead, we use the alternative to the efficient market hypothesis, the noise trader theory, to facilitate our analysis and supply context to it. By documenting the GameStop retail trading frenzy, we aim to empirically show the fundamental notions of the theory in works, using the framework of investor sentiment, irrational beliefs and limited arbitrage. Granted that the GameStop saga was primarily driven by retail investors, our thesis relates to the particular strand of literature which examines the effect of individual investors and the impact of their sentiment.

1.2 Individual Investor Sentiment, News & Social Media

Included in the section above, the paper series of De Long et al. [26, 27, 28], in general, indicate that the investor sentiment generates systematic risk likely leading to large deviations away from the fundamental values, which moreover, can sustain even in the long-term with the help of limited arbitrage. Following DSSW's conclusion regarding the

closed-end funds (CEF) pricing, Lee et al. (1991) and Brown (1999) [30, 31] empirically validate the significant relationship between investor sentiment, respectively noise trader risk, and CEF puzzle. In addition, the latter shows a significant impact on CEF's volatility. Further evidence, consistent with the noise trader theory's conclusions, is provided by Bodurtha et al. (1995) [32] suggesting that closed-end country fund premiums may be related to the U.S. market sentiment, Pontiff's (1997) proof of CEF's excess volatility or the parallel between CEF and the U.K. real estate companies by Barkham et al. (1999) [32, 33, 34]. The effect of investor sentiment using various proxies, including, but not limited to, equity share offerings, closed-fund discounts, net mutual fund redemptions or odd-lot sales, on the cross-section of stock returns is extensively documented, see Swaminathan (1996), Kothari and Shanken (1997), Neal and Wheatley (1998) or Baker and Wurgler (2000, 2006) [35, 36, 37, 38, 39]. A short overview of the literature attempting to link individual investor sentiment and the stock market is available in Baker and Wurgler (2007) [40]. Moreover, they demonstrate the predictive power of a designed six-part measure of sentiment resulting from the literature review, composed of closed-end fund discount, dividend premium, NYSE share turnover, equity share in new issues, number of IPOs, and first-day IPO returns.

The DSSW model [26, 27, 28] predicts that, as sophisticated and noise traders interact on the market, prices can largely deviate from the fundamental values with the latter trading on pseudo-signals, respectively non-fundamental news. Related, and considering the spotlight on the individual investors by the GameStop saga, we advance to sentiment measures more closely related to investor psychology. Specifically, following up is the strand of literature more intimately examining the effect of the sentiment using micro-level data, including brokerage data on individual transactions, trade imbalances, etc. In addition, a survey of literature studying retail investors' behavioural tendencies in greater detail is also included. Consistent with the noise trader theory, the presented literature ought to primarily challenge the prevailing notions on the impact of individual investors. In particular related to their ability to affect prices, coordinated behaviour among a large group of retail investors and lastly, the belief that their synchronised actions ought to be cancelled out by rational arbitrageurs and thus, negating their effect on prices. Note that, due to the voluminous literature documenting the effect of sentiment on stock markets, we are inherently selective and merely attempt to present the most relevant studies emphasising the role of individual investors and their behavioural tendencies. A brief overview of the topic can be found in Baker and Wurgler (2007) [40].

The fact that retail investors may excessively respond to pseudo-signals is backed by extensive evidence. It is established that their suboptimal respectively, excessive, trading activity, primarily caused by overconfidence, is detrimental to their expected returns,

which is at odds with the theories of rational investors, as shown by Lee (1992), Odean (1999), Barber and Odean (2000, 2001) or Chang, Hsieh and Wang (2015) [41, 4, 42, 43, 44]. Notably, Barber and Odean (2002) [5] observe retail investors' increased trading activity, particularly of speculative type, when switching to online trading in the early 1990s and hypothesise that the concomitant worsening performance is attributable to cognitive biases reinforcing overconfidence such as self-attribution or illusion of control produced by the enhanced access to technology and information.

Beyond the specific behavioural tendencies, there is substantial support for the theory that individual investors act in concert and display similar trading behaviour across different categories of stocks, as documented by Jackson (2003) [45] for Australian investors, Feng and Seasholes (2004) [46] using Chinese-based dataset and others [47]. Separately, Barber and Odean (2008) [14], using abnormal 1-day volume, previous 1-day return and news mentions as proxies for attention, demonstrate that the retail investors' buying decisions, in particular, are for the most part determined by the attention the stock has drawn on a given day. Kumar and Lee (2006) [6], in addition to identifying common tendencies of individual investors' trading activity[1], shows that retail sentiment, measured by the buy-sell imbalance, is a significant predictor of returns, particularly of firms with high retail concentration characterised by small capitalisation, high book-to-market ratio, low price and lower institutional ownership. Moreover, relating to the noise trader theory's limited arbitrage argument, higher arbitrage costs are linked to a higher sensitivity to retail sentiment. Similarly, Barber, Odean and Zhu (2009) [7] find that retail trade imbalances predict returns over the short horizon, although the effect in the longer horizons is limited to the difficult to arbitrage stocks, i.e. high idiosyncratic volatility, small capitalisation stocks, similar evidence is produced by Kaniel, Saar and Titman (2008) [48].

The above literature illustrates the effect of attention on retail investors' trading activity and the susceptibility of highly volatile stocks to sentiment, meanwhile, others have attempted to reverse the optics and found a significant positive relationship between retail traded stocks and volatility, see Foucault, Sraer and Thesmar (2011) [49], Andrade, Chang, Seasholes (2008) [50] or Brandt et al. (2010) [51].

Concluding this section related to the role of investor sentiment and behavioural tendencies on a lighter note, some of the related fringe literature is presented. Following works immerse deeper into the investor psychology and, instead of sentiment, attempt to measure mood, using both exogenous and endogenous factors to emotions. Kamstra, Kramer and Levi (2003) [52] link seasonal affective disorder (SAD) with stock market returns,

[1] Across a different group of stocks, retail investors tend to buy particular stocks when selling other and vice versa. Moreover, the behaviour is replicated independently of other investors, i.e. if one group of investors buys (sells), the other group also buys (sells).

explaining the lower returns during autumn and relatively higher returns after the winter solstice. The effect of SAD was found particularly strong for northern countries. Another discrete measure is offered by Edmans, García and Norli (2007) [53], who demonstrate that loss in major football fixtures predicts the next-day significant negative returns. Goetzmann et al. (2015) [54] find that weather-based indicators - such as cloud cover - impact institutional investors' decisions. Yet another recent study piquing interest is Edmans et al. (2021) [55] using music sentiment, derived from Spotify's 200 top songs in a given country, as a proxy of investor mood. Following the evidence from the above-mentioned literature, music sentiment was able to (positively) predict same-period returns and subsequent price reversal. Moreover, it also explained mutual fund flows and - in the absolute form - volatility as well.

1.3 News & Social Media Sentiment

So far, we have emphasised both the economic fundamentals linked and endogenous measures. Related to the latter, the following strand of literature tackles the transmission mechanism, in particular, the media of information, be it financial reports, news or social media. The expeditious development in communications technology encompassing the vast ocean of information and, more importantly, the increased access to it has led to the advancement of various methods trying to capture the signalling quality of the information, some of which are central to our thesis. The study of the impact of news is not novel, see Cutler, Poterba and Summers (1989) [56] who, however, fail to find a significant link between market aggregate moves and news coverage. On the other hand, we have Gidofalvi and Elkan's (2001) [57] simple classification algorithm predicting price movements based on news articles, or a similar attempt at predicting the impact of news stories using text classification based on SVMs by Fung, Yu and Lu (2005) [58]. Further, text mining techniques have been largely utilised to predict stock returns, turnover or volatility, ranging from assessing the similarity of news (Tetlock, 2007) [59], bag-of-words method or named entity recognition (Schumaker and Chen, 2009) [60].

Moving beyond the realm of traditional news media, alternative sources of sentiment have also spurred interest from the academic literature. Online search trends or Wikipedia page views have been shown as significant indicators of returns of individual stocks, see Da, Engelberg and Gao (2011, 2015) [61, 62], Preis, Moat and Stanley (2013) [63], or Moat et al. (2013) [64]. Yet another interesting territory, one highly relevant for our thesis, is social media. Antweiler and Frank (2004) [8] show internet message board activity, on Yahoo! Finance and Raging Bull, in particular, predict stock returns, turnover and volatility. Similarly, using popular stock trading social media Seeking Alpha, Chen et al.

(2014) [9] and Farrell et al. (2022) [11] predict future stock returns and earnings surprises. Indeed, even Reddit's *r/wallstreetbets* subreddit's so-called due diligence reports have been found to significantly predict stock returns in the past, as shown by Bradley et al. (2021) [65].

Lastly, for this section, we would like to highlight some of the interesting findings pertinent to our analysis. Firstly, the challenge of the uniform nature of investor sentiment derived from various sources. Jiao, Veiga and Walther (2020) [10] differentiate between social media and news sentiment, in particular showing that social media more resemble *echo chambers* where existing fundamental signals ceaselessly repeated, suggesting that high social media coverage is followed by periods of high volatility and turnover compared to the opposite effect of news sentiment. Secondly, consistent with the findings of Antweiler and Frank (2004) [8], Atkins, Niranjan and Gerding (2018) [15] find, using intraday data and Reuters news archive, that sentiment predicts volatility better rather than stock returns.

1.4 GameStop Related Literature

The last literature survey concerns the recent academic works exploring the GameStop saga from various angles. The examination of recent findings and hypotheses serves both for validation purposes and to establish a common base for the literature's findings and thesis goals.

In the first of a series of papers Umar et al. (2021a, 2021b, 2021c) [66, 67, 68] using a network approach examine the connection between media sentiment and high short interest stocks with an emphasis on volatility spillovers. Relevant to our analysis, the authors find a significant drop in connectedness around January 2021, hinting at non-fundamental drivers. The latter two works employ wavelet analysis and closely examine the GameStop retail trading saga. Umar et al. (2021b) [67], using Twitter and news media data in addition to short-sale volume and options data to proxy retail investor sentiment, document the significant role of sentiment on GameStop returns. The social media and news data, however, only consist of publication counts, whereas, instead of proxies, we mine sentiment directly from the Reddit posts from where the trading frenzy originated. Last, in the paper series, the wavelets framework is utilised to model comovements between the heavily shorted stock indices with particular emphasis on GameStop. While Umar et al. (2021c) [68] document strong comovement between GameStop and high short interest indices, more detailed analysis concerning financial contagion, in particular spillovers to particular assets, is not offered. Similarly, Long, Lucey and Yarovaya (2021) [69] using scraped comments from the r/wallstreetbets subreddit find an effect,

albeit quite weak, of sentiment on GME returns, sampled at 1-min intervals, including an ambiguous bidirectional causal linkage. Moreover, the sentiment, complicatedly further categorised by six emotions, shows extremely low comovement with the stocks' returns.

The strand of literature focusing on the role of individual investors within the GameStop saga extensively utilised options data to describe the effect of investor sentiment. Fusari, Jarrow and Lamichhane (2020) [70] utilise options data to identify asset market bubbles, using the GameStop bubble at the start of 2021 as an example. On a different note, Jones, Reed and Waller (2021) [71] examine the restrictions imposed on trading and increased margin requirements on specific stocks, showing that the effect of trading restrictions, different from the effect of mentions on Reddit, led to sizeable price falls with no subsequent recovery. The increased activity in the options market which followed is demonstrated to accommodate large transfers from option buyers to option writers due to larger prices, open interest and greater implied volatility.

The role of options during the GameStop retail trading saga is further highlighted by Allen et al. (2021) [72], where it is showcased that options were used to circumvent the trading and short-sale constraints. Further, Baltussen (2021) [73] attempts to explain the extreme dynamics in GME stock and increased activity on options markets by institutional investors' need for so-called *gamma hedging*.

2 Methodology

2.1 Text processing

Using the Reddit API service, we obtain data on the users' submissions on the *r/wallstreetbets* subreddit which besides including all identifying features such as username, number of comments, score and other metadata, are also timestamped. For our purposes and to greatly reduce the memory strain, we only retain submissions' title, text body, id and the UTC timestamp in our sample. Additionally, we require that the posts are not removed, respectively deleted, by the administrators or users.

Firstly, we clean the text using the typical procedure ridding the data of formatting characters specific to Reddit posts, URL addresses, numbers, stopwords and other irrelevant traits. One of the identifying features of social media data is the usage of emojis, which in our case is quite abundant. We explicitly deal with this notion in two ways yielding two different datasets. Using the emojis' Unicode we identify them in the text and subsequently either remove them or translate them into text using the appropriate CLDR Unicode name. Provided that the usage of emojis is quite prevalent and one of the identifying features of social media, we believe that their usage possesses relevant information. Thus, we opt for the latter option of translating them. In a similar vein, sentiment analysis tools often yield scores even for emojis.

Following the cleaning procedure, we proceed with the tokenising, stemming and lemmatising. Moreover, we extract bigrams as they provide greater context and capture some of the semantics omitted when obtaining only word tokens. Expanding further to trigrams, however, does not yield more substantial information. Notably, the bigrams aptly capture a greater range of semantic information as emojis' CLDR Unicodes often contain multiple words.

2.2 Sentiment extraction

The principal tool of our analysis used to extract sentiment from the gathered textual data is the *Valence Aware Dictionary and sEntiment Reasoner* (VADER) [74]. The tool relies on lexicon and rule-based sentiment analysis and is specifically accommodated for dealing with social media data. Given that VADER can properly score even complex structured sentences, including negations, contractions, punctuation, capitalisation, degree modifiers, slang, emojis and others, we apply the sentiment analysis to non-cleaned text as some of the text cleaning procedures can lead to the loss of semantic value otherwise retained. VADER yields four scores pos, neu, neg and $compound$. The former three signify the proportion of words in a given text which are scored positive, neutral

or negative on a lexicon basis. Hence, they do not reflect the rule-based approach, i.e. context. Instead, they represent the independent distribution of words' classifications. The most important measure for us and mostly utilised in our analysis is the *compound* score, which essentially constitutes a normalised, weighted composite sentiment score taking values between -1 (extremely negative) and 1 (extremely positive).[2]. Establishing thresholds typically used in the literature allows us to transition between the scoring and classification problems. *Compound* score higher (lower) than 0.05 (-0.05) is identified as positive (negative) sentiment. While scores smaller than 0.05 in absolute value are considered neutral.

One of the key elements of sentiment analysis is the used lexicon containing the words and the assigned scores. While the VADER sentiment lexicon is considered the golden standard, moreover, it is able to assess both polarity and intensity of sentiment, especially in relation to social media, we are compelled to adjust the lexicon to better reflect upon the very specific nature of the language used on the *r/wallstreetbets* subreddit. Therefore, following the assessment of the most used words and their corresponding scores from the lexicon, we adjust them based on the perceived bullish or bearish signal. We limit the number of changes made, so as to not disrupt the original lexicon. Most of the changes were made due to the specific WSB language, financial terminology or emojis. For instance, we established scores for terms like *buy, short, long* or *short* and some translated emojis such as *rocket, fire* or *moon* which reflect bullish signals. For more information regarding the VADER sentiment analysis tool and the construction of the lexicon refer to Hutto and Gilbert (2014) [74].

2.3 Wavelet analysis

2.3.1 Applied literature

Before we proceed with the description of the method, we briefly present some of the key literature, including the applications of the wavelet analysis to procure some context for our application of the wavelet analysis. Wavelets have found their use across a vast number of disciplines ranging from meteorology, oceanography, and geology to signal and image processing, see Kumar and Foufoula-Georgiou (1997) [75] or Kronland-Martinet (1988) [76]). The fields of economics and finance have been remarkably fairly late additions to the long list of disciplines utilising wavelet methods. Some interesting applications include Ramsey and Zhang's (1996) study of foreign exchange data structure, and Ramsey and Lampart's (1998) [77] examination of the money-income relationship.

[2]A more detailed description is given in the paper of Hutto and Gilbert (2014) [74] or on the GitHub repository containing the vaderSentiment package for Python available here: `https://github.com/cjhutto/vaderSentiment`

For an overview of wavelets method prominence in the field of economics refer to Crowley (2007) [78]. Regarding the finance-related works, wavelet analysis has been leveraged predominantly in studying stock comovements (Rua and Nunes, 2009) [79] and financial contagion (Gallegati, 2012) [80]. The key advantage of the wavelets framework is its non-parametric approach to analysing dependency in both time and frequency spaces. While the frequency dimension essentially grants us a pathway to examine short- and long-run dependencies, the time localisation ensures keeping information on how the relationship evolves through time. For a more detailed description of the applied methods, including their applications, we refer the reader to the seminal work of Torrence and Compo (1998) [81], or previously mentioned Crowley (2007) [78] for explanation within the economics realm, and others (Kumar and Foufoula-Georgiou, 1997; Grinsted et al., 2004; Rua and Nunes, 2009) [75, 82, 79].

The key tool applied in our work is the wavelet framework owing to its mentioned ability to localise in both time and frequency domains. Moreover, unlike other mathematical methods parsing through the frequency domains such as Fourier analysis, the wavelet analysis does not assume time-stationarity in the process. Therefore, the tool combines the two worlds by examining the dependence of two series in the time-frequency space. Its non-parametric nature also presents an advantage. Let us present a brief overview of the wavelet analysis framework, particularly employed in the study, including continuous wavelet transform (CWT) cross wavelet transform (XWT), wavelet coherence (WTC) and phase difference.

A wavelet function, by definition, must have a zero mean and be localised in both time and frequency domains. In general, it can take the following form:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)$$

The term $\frac{1}{\sqrt{s}}$ ensures the wavelet's unit variance. The wavelet is characterised by two parameters, u around which it is centred, represents location, while s defines the scale or dilation of the wavelet. In simplified terms, scale is particularly important with respect to how a signal *stretches* over time, higher (lower) scale is consistent with a wider (narrower) time scale. Note that, the relationship between scale and frequency is inverse. The unilateral shift of the wavelet in time is described by the shift of u. The admissibility condition of a wavelet, i.e. the need to be localised in both time and frequency space, is described as

$$\int_0^\infty \frac{|\Psi(t)|^2}{t} df < \infty\,,$$

where $\Psi(t)$ represents the Fourier transform of $\psi(t)$. The condition ensures that the Fourier transform of $\psi(t)$ is zero at zero frequency $|\Psi(t)|^2\big|_{t=0} = 0$ and that the wavelet has a zero mean $\int_{-\infty}^{\infty} \psi(t)dt = 0$, i.e. it oscillates. For our purposes we resort to using the *Morlet* wavelet due to its favourable properties in both time and frequency localisation. The wavelet is complex and defined as

$$\psi_0(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-\frac{1}{2}t^2} ,$$

where ω_0 represents the non-dimensional frequency and is generally set at 6.

2.3.2 The continuous wavelet transform

Consider a time series $x(t)$, then for a scale and location parameters s and u, the continuous wavelet transform (CWT) $W^X(u,s)$ is defined as the convolution of a time series $x(t)$ with the normalised and scaled wavelet $\psi(.)$, i.e.

$$W^X(u,s) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt ,$$

where $*$ denotes a complex conjugate. By varying s and t, parameters, respectively stretching the wavelet over time and translating it along the time axis, we can describe how amplitude changes over time as well as the phase, indicated by the complex arguments. Effectively, we obtain the CWT coefficients for different values of s and t which represent a set of basis wavelet functions $\psi_{u,s}(t)$ into which the original series can be decomposed. These wavelets come from the specified *mother wavelet*, in our case Morlet ($\omega_0 = 6$). The wavelet power is defined as $|W^X(u,s)|^2$, this attribute plays a principal role in assessing the dependence of the two series.

2.3.3 Wavelet coherence

First, consider two time series x_t and y_t, then their cross wavelet transform (XWT) is defined as $W^{XY} = W^X W^{Y*}$ and the corresponding cross wavelet power as $|W^{XY}|^2$, as defined by Torrence and Compo (1998) [81]. W^X and W^Y represent the continuous wavelet transforms of the time series. The resulting cross wavelet power locates time periods in which both series exhibit high common power. Meanwhile, the dependence of the time series in the time-frequency dimensions, while not necessarily sharing high common power, is given by the concept of wavelet coherence. Following Torrence and Compo (1998) [81] the wavelet coherence is defined as

$$R^2(u,s) = \frac{|S\big(s^{-1}W^{XY}(u,s)\big)|^2}{S\big(s^{-1}|W^X(u,s)|^2\big) S\big(s^{-1}|W^Y(u,s)|^2\big)} ,$$

where S is a smoothing operator over time and scale.[3] The squared wavelet coefficient takes values between 0 and 1, it can thus be interpreted as a sort of time-frequency variant of a traditional correlation coefficient. Wavelet squared coherence closer to one (zero) indicates a higher (lower) dependence of the two series. The procedure for testing statistical significance is given by the Monte Carlo estimation methods which, in brief, involve generating a large set of surrogate data with the same $AR(1)$ coefficients and computing their wavelet coherence for comparison. For a detailed description of the procedure refer to Torrence and Compo (1998) and Grinsted et al.(2004) [81, 82].

2.3.4 Phase

Note that wavelet coherence while being able to describe the degree of covariation, it does contain information concerning the positive and negative dependence. To examine that, it is advocated to use the information carried by the imaginary part of the WCT, the phase difference which signifies the lead and lag relationship between the two series and can be indicative of causality. The phase difference is defined as follows

$$\phi^{XY}(u,s) = tan^{-1}\left(\frac{\Im\{S\big(s^{-1}W^{XY}(u,s)\big)\}}{\Re\{S\big(s^{-1}W^{XY}(u,s)\big)\}}\right),$$

where $\Re$ and $\Im$ represent the real and imaginary parts of the cross wavelet transform. For instance, a phase difference of 45° indicates that the first series leads the second by 45° and that they co-move together. It should be mentioned that phase difference while possessing important information about the relationship of the two series, is only indicative of causality and thus, shall be interpreted with caution.

Lastly, similarly to other filtering methods, the wavelet transform suffers from edge effects as the wavelet is not completely localised in time because of the finite length of the data. Particular caution needs to be taken in the areas where discontinuities occur due to edge effects which cannot be ignored. These are grouped under the term *Cone of Influence* (COI), see Grinsted et al. (2004) [82].

2.4 Vector error correction model

As discussed above, while the wavelet analysis can reveal bivariate dependence parsing through both time and frequency space and further detect the change in the dependence relationship throughout time, it lacks in respect to determining causality. Therefore, in this aspect, we also resort to expanding our toolkit with a traditional time-series analysis in the form of vector correction models. The main advantage of which, is that it allows us to statistically test and infer causality both in the short- and long run. The latter is

[3] For a more detailed description refer to Grinsted et al. (2004) [82]

related to the concept of cointegration. The extension of our toolkit will be primarily motivated by the results of the wavelet analysis. Pursuing this, we look to comment on the causality between sentiment and GME, and other relevant stocks. Moreover, the below-introduced model apparatus will provide us with formal statistical testing of the comovements concerning the spillovers from GME.

In the following section, we introduce the concept of cointegration and vector error correction models. For brevity, we constrain the description only to the main aspects of the model and omit the theoretical background and the mathematics of the model apparatus. Furthermore, regarding the choice of model specification and estimation procedure, we also skip a detailed description. Concerning these technicals, when pertinent, the reader is referred to the relevant literature.

The definition of cointegration carries two formal requirements. In the bivariate case, the two time series are cointegrated if they are of the same order of integration $I(d)$ and there exists a linear combination of the two that is integrated of order $I(d-b), b > 0$.[83] Essentially, using the concept of cointegration we aim to examine whether a linear combination of non-stationary series can be stationary, for instance, a linear combination of non-stationary stock price series exhibiting a stationary relationship. Through the lens of cointegration analysis, we can study the long-term relationship between the variables in levels such as sentiment and stock prices. Further, utilising an error-correction model dynamics of the adjustment toward the long-run equilibria as well as short-run dynamics can be examined. Ultimately, following the precursory role of the wavelet analysis, we further scrutinise the causality between the sentiment and the relevant stocks in the subreddit while making a distinction between short- and long-run dynamics, which ought to be highlighted in the wavelet framework.

Let us now ensue with the description of the model. We utilise the vector error correction model (VECM) in the following Hendry and Juselius' (2001) [84] form

$$\Delta X_t = \sum_{j=1}^{k} \Gamma_j \Delta X_{t-j} + \Pi X_{t-1} + \alpha\mu + \gamma + \alpha\rho t + \tau t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{IN}_N[0, \Omega_\epsilon], \qquad (1)$$

where Γ_j are $N \times N$ coefficient matrices of the vector autoregressive (VAR) components with up to k lags and ϵ_t is a vector of N unobserved error terms. The $\Pi(N \times r)$ matrix contains the error correction coefficients and it can be rewritten as $\Pi = \alpha\beta'$, where $\beta(N \times r)$ represents the matrix of cointegrating vectors while $\alpha(N \times r)$ contains the loading coefficients describing the speed of adjustment to the equilibrium. Note that Π matrix has reduced rank, i.e. $r < N$, r denotes the number of cointegrating relationships between the variables. Were $r = 0$, then Π has a zero rank and we have no cointegration in the system,

on the other hand, were Π to have full rank it would mean that our dependent variables are all stationary and VECM reduces to a VAR model with level variables. Finally, the deterministic terms contained in the equation can also be rewritten as $\pi = \alpha\mu + \gamma$ and $\delta = \alpha\rho + \tau$. With the inclusion of the deterministic terms, there are five possible cases, the description of which, however, is omitted here, for a detailed description please refer to Hendry and Juselius (2001)[84]. Regarding the choice of correct specification of the model we closely adhere to the procedure set out in Filip et al. (2019) [85].

Concerning the determination of the rank of the Π matrix, i.e. the number of cointegrating relationships, we follow the Johansen test [86] which is based on testing the number of non-zero eigenvalues of the said matrix. Specifically, there are two tests based on trace and maximum eigenvalue statistics. The former tests the null hypothesis of r cointegrating relationships against the alternative of at least $r + 1$ cointegrating relationships, the trace statistic is defined as follows

$$LR_{trace} = -T \cdot \sum_{i=r+1}^{N} \ln(1 - \hat{\lambda}_i) \tag{2}$$

The slightly different maximum eigenvalue then tests against a specific alternative hypothesis of $r + 1$ cointegrating relationships and the test statistic is given as

$$LR_{max} = -T \cdot \ln(1 - \hat{\lambda}_{r+1}) \tag{3}$$

A deeper examination of the tests and their intuition is beyond the scope of our work, for that see Johansen (1991, 1995) [86, 87]. Following the estimation of the fully specified model, we are able to test both the short- and long-run dynamics between the variables via the standard procedure of Granger causality testing, which we shortly cover below.

Granger causality is based on a relatively simple premise saying that if past values of X bring additional value to predicting Y while having already accounted for the effect of past values of Y then X Granger causes Y. In a bivariate example, this can be formally tested as follows

$$\Delta x_t = \alpha_{10} + \sum_{i=1}^{k} \alpha_{1i} \delta x_{t-i} + \sum_{j=1}^{k} \beta_{1j} \Delta y_{t-j} + \epsilon_t \tag{4}$$

$$\Delta y_t = \alpha_{20} + \sum_{i=1}^{k} \alpha_{2i} \delta y_{t-i} + \sum_{j=1}^{k} \beta_{2j} \Delta x_{t-j} + \nu_t \,, \tag{5}$$

where testing the null hypothesis that X does not Granger-causes Y boils down to testing the joint significance of the relevant variables as follows

$$H_0 : \beta_{21} = \beta_{22} = ... = \beta_{2k} = 0$$

In general, for a system of N endogenous variables and k order of lags, there can be $N!$ Granger causality tests with N^2 possible results. For our purposes, the short-run dynamics are tested via the above-specified approach while the long-term dynamics rely on the error correction terms contained in the $\Pi(N \times r)$ matrix.

3 Data

3.1 Textual and sentiment data

Central to the thesis is the social media data manually scraped from Reddit, in particular from the subreddit most involved in the GameStop saga, the *r/wallstreetbets*. Using Reddit's API service, via wrappers PRAW and PSAW, we access data on the number of posts as far back as January 2021.[4] Considering the principal role of the *r/wallstreetbets* subreddit, we retrieve data on specific posts corresponding to a set of GME related keywords, including relevant discussion megathreads and daily discussion threads posted for each preceding and following trading day. This results in 456,055 unique *r/wallstreetbets* posts, after removing duplicates, user or moderator deleted submissions, spanning from January 2021 until December 2021, including the submissions' title, body and other identifying features including a timestamp with accuracy up to a second. The scraped posts include 59 *megathreads* created to moderate discussion during the most active periods and 419 periodical discussion threads which are posted each day before and after the market opens, where users discuss contemporaneous sentiment and strategy. Using these threads, we track the general discussion and assess the relevance of GME within the subreddit by extracting up to 1000 comments and searching for GME-related keywords, see Fig. 1.

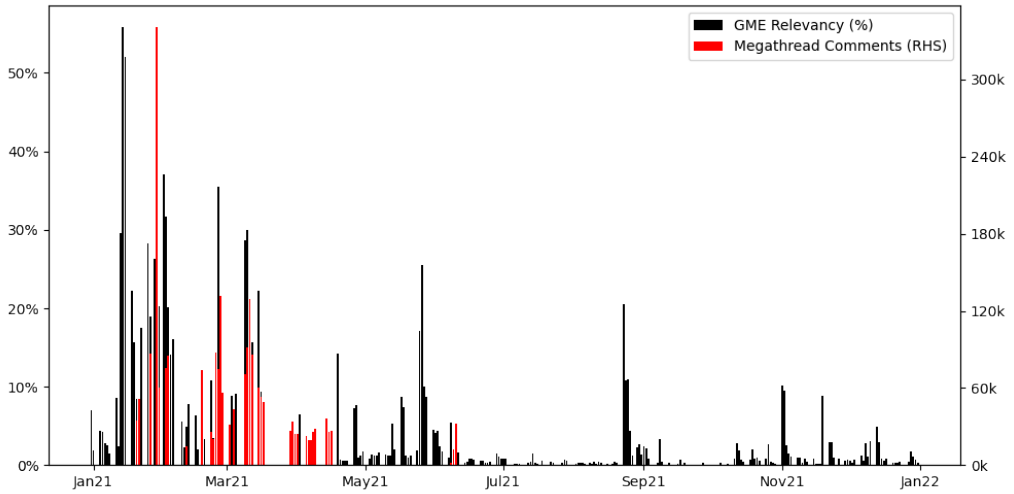
To obtain the sentiment measure, the submissions' titles are analysed. Resulting in more than 3 million word tokens which are used for the textual analysis, while for the sentiment analysis soft cleaning procedure is employed to remove web addresses, Reddit-specific formatting and other non-informational features. However, note that word-shape, punctuation, or UTF-8 encoded emojis are not removed, as they all possess valuable information for extracting sentiment. We use the terms Reddit sentiment and retail investor sentiment interchangeably when referring to the sentiment measure ($Sent$). The Reddit sentiment is aggregated as a sum of compound sentiment scores according to the desired frequency, to capture the intensity of sentiment as well as the mood, i.e. higher $Sent$ will reflect both higher activity and heightened mood of the discussion. For a more detailed description of the data, small samples of the scraped submissions areincluded in Tab. 1 and 2 below. These submissions constitute the ten most upvoted and commented *r/wallstreetbets* posts from January 2021.

While the sentiment is continuously collected throughout all 24 hours, the market adheres to a different schedule and given that we are interested in the interactions between

[4]PRAW and PSAW documentations are available here `https://praw.readthedocs.io/en/stable/`, `https://psaw.readthedocs.io/en/latest/`

Fig. 1: GME relevancy

GME Relevancy shows the percentage of mentions in 1000 sampled comments from each daily discussion thread, i.e. 50% around February denotes that from the 1000 sampled comments from the daily thread 50% had mentioned GME or contained a related keyword. To further highlight most intense activity periods, we include number of comments on megathreads (RHS), which were created to contain the discussion surrounding GME.

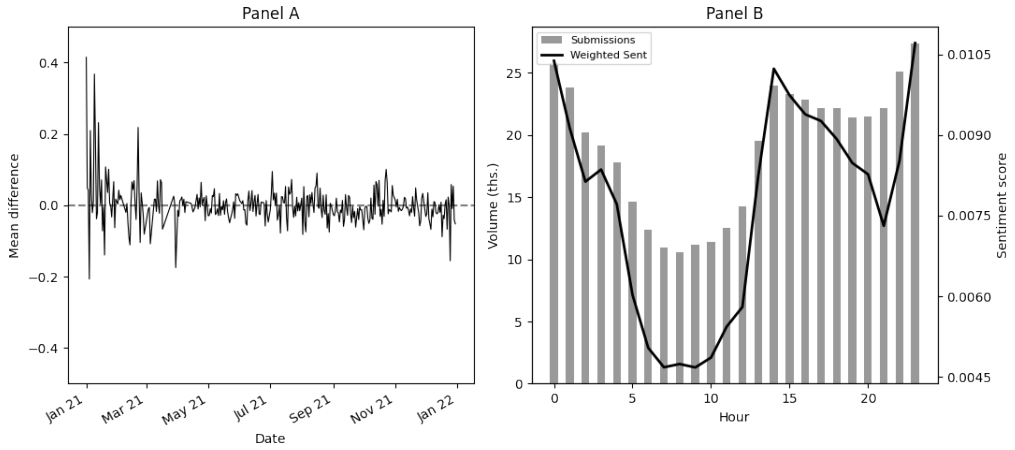


the scraped sentiment and intraday financial time series, we obliged to discard the sentiment data outside of the trading hours (days) accordingly. Nevertheless, we believe that the subset data remains equally informative even in the case of intraday 30-min sampling frequency. Fig. 2 shows that there is not a huge discrepancy between the sentiment on- and off-trading hours. Further, the distribution of sentiment throughout the trading day shows that it culminates before the open and dwindles during the trading hours before rising once more after the closing bell. However, as can be observed, the biggest volume of submissions is contained within the trading hours and a few hours after the end of the trading, further evidencing that the subset still captures most of the sentiment.

Let us briefly discuss one of the limitations of our study related to the scraped dataset. By the rules of the subreddit, all submissions must possess a descriptive title. Often, we are not able to extract all of the posts' content beyond the title and other identifying information. Posts may be in the form of images, videos, links, etc., which, however, still contain a title. This feature is reflected in a significantly lower amount of posts which contain a body of text. One possible proxy measure of a submission's sentiment, respectively mood within the posts, could be given by mining the submission's comments. But the number of comments on each post varies significantly, which could present a bias in the measure. Nevertheless, we believe that our reliance on the submissions' title is justifiable as it provides the most context without introducing unnecessary noise. Therefore, the sentiment measure is derived from the posts' titles given their explanatory power.

Fig. 2: Difference in sentiment off- and on- trading hours

Figure on the left illustrates the mean difference between the means of daily aggregated sentiment scores on- and off-trading hours. Positive deviation indicates that sentiment off-trading hours was higher and vice versa. The right hand-side plot shows the volume weighted mean sentiment throughout the day, i.e. both the intensity and overall mood of discussion are reflected. The hours on the x-axis correspond to CET + 1 time thus, the trading hours of the US stock exchange correspond to 16:30 - 23:00 CET + 1. It is shown that sentiment mostly builds up before the open and loses some of the traction throughout the trading before again gaining after the close.



Separately from the manually scraped data, we also employ the news sentiment series from Bloomberg for GME, AMC, BB and CLOV, in addition to the Twitter sentiment specifically for GME. The Twitter series will serve mainly two functions. Firstly, to validate our scraped dataset and the mined sentiment. Secondly, with respect to comparing news sentiment, it may offer a distinctly different point of view provided by other social media. It should be noted, however, that while these datasets are readily usable, we possess limited knowledge regarding the functioning of the sentiment extraction tools and the sentiment scoring as opposed to being able to tweak the sentiment extraction to accommodate the specific lexicon and language on the subreddit. Therefore, the comparison between the datasets and our scraped sentiment might be rendered weaker. The news sentiment is aggregated daily and contains the news publication counts. The individual news or *stories* are then categorically labelled positive, negative or neutral. Hence, opposed to our scoring approach, it only offers a classification of positive and negative *stories*. The same methodology is applied to the Twitter series.[5] The summary of the data is given in Tab. 3.

[5]For more details on the construction of the sentiment series see Bloomberg's whitepaper `https://www.bloomberg.com/professional/sentiment-analysis-white-papers/`

Tab. 1: Sample Reddit Posts from January 2021

Sample of the most upvoted and commented posts from January 2021. Some of the identifying features are presented for demonstration, including the submissions' title, score, hash-id, subreddit, number of comments, submissions text body (if applicable) and UTC timestamp.

Most Upvoted Posts

title	score	id	subreddit	num_comments	body	created	datetime
GME YOLO update — Jan 22 2021	91199	l2x7he	wallstreetbets	6332		1.611379e+09	2021-01-23 05:14:08
Honorary WSB Autist award goes to Chamath Pali...	82970	l69jz5	wallstreetbets	2696		1.611797e+09	2021-01-28 01:23:12
GME YOLO update — Jan 25 2021	81959	l4xje1	wallstreetbets	5514		1.611638e+09	2021-01-26 05:05:36
Can I get a flair for buying GME at the litera...	58019	l4sg3u	wallstreetbets	4153		1.611624e+09	2021-01-26 01:18:21
GME YOLO update — Jan 13 2021	52148	kwpviw	wallstreetbets	3188		1.610601e+09	2021-01-14 05:04:19
$500 Donation For Every $50 Increase in GME Pr...	46845	l5be2b	wallstreetbets	1132	[deleted]	1.611686e+09	2021-01-26 18:29:54
GME YOLO update — Jan 19 2021	39183	l0t8ng	wallstreetbets	1615		1.611119e+09	2021-01-20 05:05:55
TRUTH about GME effect!	38375	l67c0p	wallstreetbets	1903		1.611791e+09	2021-01-27 23:51:36
The GME Thread, Part 3.14, for January 27, 2021	36694	l6cb1x	wallstreetbets	49327	Reddit's Engineering team has asked us to rota...	1.611804e+09	2021-01-28 03:26:08
I want to thank you guys for saving my best fr...	36277	l667hb	wallstreetbets	1233	Monday afternoon I took my best friend, an Ame...	1.611788e+09	2021-01-27 23:00:39

Most Commented Posts

title	score	id	subreddit	num_comments	body	created	datetime
The GME Afterhours Thread: Part 4.20 on 27 Jan...	27604	l6er79	wallstreetbets	94621	Stop spamming copy pastes you boomers. Instaba...	1.611812e+09	2021-01-28 05:26:35
The GME Thread Part 1 for January 26, 2021	14456	l5c0nr	wallstreetbets	93889	Good luck today. [Here's some WSB stats.](http...	1.611689e+09	2021-01-26 19:16:56
GME Thoughts, YOLOs, Gains, Stonk Updates,...	18418	l4lmrx	wallstreetbets	93756	Thanks all for the quick rise to max comments ...	1.611601e+09	2021-01-25 18:56:35
GMREEEEEEEEEEEE Containment Thread - GME shitp...	14770	l2ljpt	wallstreetbets	93388	Don't be doxxing citron or anyone else. That's...	1.611342e+09	2021-01-22 19:07:41
GME Megathread Part 2	14249	l4syrd	wallstreetbets	86749	Keep all $GME discussion and memes in here. No...	1.611625e+09	2021-01-26 01:41:39
The GME Thread, Part 2.1, for January 27, 2021	14039	l692dj	wallstreetbets	72294	New thread requested due to the pure chaos our...	1.611796e+09	2021-01-28 01:03:50
GME Megathread - Lemon Party 2: Electric Boogaloo	9072	l1xtan	wallstreetbets	51390	No inauguration today so Citron *may* be able ...	1.611262e+09	2021-01-21 20:49:03
The GME Thread, Part 3.14, for January 27, 2021	36694	l6cb1x	wallstreetbets	49327	Reddit's Engineering team has asked us to rota...	1.611804e+09	2021-01-28 03:26:08
GME Thread: The Wreckoning	12034	l0hhqg	wallstreetbets	46765	Post all your GME hopes, prayers, and stupidit...	1.611082e+09	2021-01-19 18:47:16
GME Megathread - Lemon Party (keep your shitpo...	11861	l19k9r	wallstreetbets	34626	[deleted]	1.611180e+09	2021-01-20 21:52:13

Tab. 2: Scored sample of submissions

neg, *neu* and *pos* correspond to the proportion of scored as negative, neutral or positive in the given sentence. *Compound* score represents a normalised, weighted composite sentiment score taking values from -1 to 1.

itle	neg	neu	pos	compound
$GME at Premarket 01/25	0.000	1.000	0.000	0.0000
Ryan Cohen's Roller Coaster rocket rocket rocket rocket	0.000	0.182	0.818	0.9517
GME Megathread - Lemon Party 2: Electric Boogaloo	0.000	0.722	0.278	0.4019
"If you sell GME, you're a pussy." - Warren Buffet probably	0.219	0.781	0.000	-0.4215
Thank you Reddit and WSB! I can pay off my student debt! Let's do this! GME Gang BB gang	0.205	0.687	0.107	-0.3129
There's gonna be hella GME IV crush in the coming day + GME price target(based on technicals)	0.104	0.896	0.000	-0.2177
NAKD went up more than 100% today! This stock will fly over 10$ in a couple of days. Make it another GME maybe?rocket rocket	0.000	0.712	0.288	0.8516
SELL AMC, SELL BB, SELL NOK, SELL EXPR. BUY BUY BUY GME	0.455	0.203	0.341	-0.4215
Available shares on iBorrowdesk For GME at an all time low! rocket rocket rocket	0.082	0.352	0.566	0.9080
GME up 28% in premarket rocket rocket	0.000	0.357	0.643	0.8402

In order to better compare the scraped sentiment and the Bloomberg sentiment series, we transform the former to reflect counts, according to the convention dictating that a compound score higher (lower) than 0.05 (-0.05) is counted toward positive (negative) counts. Values lower than 0.05 in absolute value are considered neutral in sentiment and accordingly discarded for the purpose of comparison. While summary statistics offer some comparison, we also include Fig. 3 in the Appendix which portrays the differences between the data. Ignoring the huge disparity in the dataset sizes, there is a stark difference between news and social media. The former is shown, notably, volatile with many reversals swinging from one extreme to another. Comparatively more stable are both social media sentiments, although, we observe a large difference between the Reddit and Twitter sentiments. While *Sent* shows a much stronger bias toward positive sentiment, Twitter users do not exhibit such behaviour. Instead, tweets are predominantly negative throughout, i.e. more than 50% of tweets were considered negative. Moreover, during the most active period at the beginning of the year, sentiment on Twitter stayed deeply negative compared to high optimism on the subreddit. Nonetheless, this emphasises the unique point of view and contribution of the Reddit retail investor sentiment gauge, which provides distinct optics compared to news and Twitter sentiment. Further, we continue to stress the direct connection between the GameStop retail trading frenzy and Reddit, where it originated.

3.2 Financial data

The historical stock prices are rerieved from several sources based on the frequency. Daily Open-High-Low-Close (OHLC) stock quotes and volume data for the six sampled tickers (GME, AMC, BB, NOK, CLOV and RKT) starting from January 2020 are obtained from Yahoo Finance. The sample choice is primarily given by the results of the textual analysis, specifically, the mining of the most mentioned tickers alongside GME in the queried posts. Thus, the sample consists of most frequently mentioned tickers by posters and commenters along GME in the related threads. A particular focus in this selection is given to the beginning period of 2021. In addition to the individual stocks, we also obtain short interest index (Citi US Short Interest Equity Index) and broad market index from Bloomberg.

Higher frequency, intraday OHLC data based on 30-minute intervals are likewise downloaded from Bloomberg with the sample period spanning one year, beginning November 2020. Furthermore, we also obtain the implied volatility series from Bloomberg for GME. The difference between the historical and implied volatility could offer great insight into the sentiment implied by speculative trading. Moreover, given the importance of options

Tab. 3: News and Twitter sentiment

Descriptive statistics of Reddit sentiment, Bloomberg news and Twitter sentiment data sets. The Pos and Neg scores for our measure describe the proportion of words scored positively or negatively. Comp score corresponds to the normalised, weighted sentiment score. Both news and Twitter sentiment series contain the counts of news stories or tweets, scored either positive or negative, according to whether it sends bullish or bearish signal to the investor. Count represents the number of assessed news stories or tweets, Pos (Neg) stand for positively (negatively) scored stories. Notice the difference between average counts and data sizes.

	GME (Reddit)			GME (News)			GME (Twitter)		
	Neg	Pos	Comp	Count	Pos	Neg	Count	Pos	Neg
N	456054	456054	456054	521.0	521.0	519.0	521.0	508.0	514.0
Mean	0.06	0.17	0.18	143.7	6.8	−8.5	2312.3	146.1	−183.6
Std	0.13	0.22	0.41	423.4	22.1	34.4	6814.5	420.1	568.4
Min	0.00	0.00	−0.99	1.0	0.0	−438.0	17.0	1.0	−5960.0
25%	0.00	0.00	0.00	13.0	0.0	−3.0	104.0	7.0	−116.8
50%	0.00	0.00	0.00	36.0	1.0	−1.0	609.0	39.0	−47.0
75%	0.03	0.30	0.49	106.0	4.0	0.0	1751.0	103.0	−10.0
Max	1.00	1.00	1.00	4587.0	235.0	0.0	74211.0	4233.0	−1.0
Kurtosis	8.77	0.76	-0.45	61.9	61.5	83.0	59.3	49.0	52.4
Skewness	2.77	1.22	0.03	7.3	7.2	−8.4	7.1	6.5	−6.8
cont.		AMC			BB			CLOV	
	Count	Pos	Neg	Count	Pos	Neg	Count	Pos	Neg
N	521.0	521.0	521.0	511.0	517.0	521.0	338.0	255.0	244.0
Mean	138.9	5.9	−6.2	27.9	1.4	−1.4	21.4	1.5	−2.1
Std	246.6	17.7	16.7	43.0	4.8	4.4	44.8	8.5	4.6
Min	3.0	0.0	−200.0	1.0	0.0	−44.0	1.0	0.0	−40.0
25%	39.0	0.0	−5.0	7.0	0.0	−1.0	4.0	0.0	−2.0
50%	74.0	1.0	−2.0	13.0	0.0	0.0	10.5	0.0	0.0
75%	140.0	4.0	0.0	30.0	1.0	0.0	25.0	0.0	0.0
Max	2631.0	261.0	0.0	354.0	48.0	0.0	545.0	118.0	0.0
Kurtosis	41.1	96.8	75.9	19.2	44.9	42.0	89.0	146.1	28.4
Skewness	5.7	8.3	−7.7	3.9	6.2	−6.0	8.5	11.5	−4.6

data, as mentioned earlier, implied volatility may contribute a great deal to our understanding of the saga. In our analysis, we predominantly use log-transformed prices and log returns because of the better properties of their distribution. Additionally, for the wavelet analysis, the data are standardised to suit better the distribution assumptions of the framework. Moreover, in some cases, the data are also detrended, particularly the volatility series. As for our volatility gauge, we employ simple historical volatility computed on a rolling 10-intervals basis depending on the sampling frequency. The summary statistics for the daily and intraday 30-min returns and volatility series are found in Tab. 4 and Tab. 5, respectively.

Summarising the data at our disposal, in short, using the scraped Reddit data containing nearly half a million entries, we extract our endogenous sentiment measure from the submissions' titles. In addition to it, we obtain daily aggregated Bloomberg news and Twitter sentiment indices for our sample. Utilising the timestamped Reddit sentiment, aggregated on the according frequency, and the financial data on returns and volatility from January 2021 to November 2021, we run the bivariate wavelet analyses. Meanwhile, the news sentiment from Bloomberg covers the period from December 2019 until Decem-

Tab. 4: Return Series Data

Descriptive statistics of the stock-level data including GME, AMC, BB, NOK, CLOV, RKT and Citi short interest equity index CIEQUSSI.[6] The series are log-transformed thus, the computed returns represent the corresponding log-differences. Intraday data is based on 30-min intervals and correspond to period from November 2020 to November 2021, while daily data is collected from December 2019 until December 2021. Number of observations, mean, standard deviation, quantiles including minimum and maximum, kurtosis and skewness are presented. Jarque-Bera (J-B) test statistic is also included, *,** ,*** denote the 10%, 5% and 1% significance levels, respectively.

	GME	AMC	BB	NOK	CLOV	RKT	CIEQUSSI
Log-prices daily							
N	512	512	512	512	389	351	512
Mean	1.960	2.218	1.971	1.486	2.257	2.891	4.882
Std	1.693	1.011	0.404	0.194	0.261	0.142	0.048
Min	-0.357	0.683	1.065	0.880	1.396	2.634	4.710
25%	0.174	1.434	1.581	1.370	2.087	2.780	4.859
50%	1.401	1.905	1.954	1.432	2.303	2.884	4.893
75%	3.796	3.460	2.319	1.677	2.402	2.985	4.914
Max	4.464	4.136	3.223	1.875	3.098	3.609	4.981
Kurtosis	-1.790	-1.132	-1.098	-0.342	1.031	1.384	0.661
Skewness	0.086	0.517	0.040	-0.107	-0.458	0.730	-0.857
J-B stat	67.4928***	49.2971***	25.5173***	3.5387	29.1250***	56.1168***	69.3143***
Daily Data							
Mean	0.013	0.010	0.002	0.002	-0.000	0.001	0.000
Std	0.130	0.168	0.055	0.035	0.067	0.055	0.006
Min	−0.600	−0.566	−0.416	−0.284	−0.236	−0.327	−0.041
25%	−0.031	−0.040	−0.022	−0.011	−0.024	−0.017	−0.003
50%	−0.001	−0.007	−0.002	0.000	−0.001	−0.001	0.000
75%	0.034	0.037	0.021	0.014	0.013	0.012	0.003
Max	1.348	3.012	0.327	0.385	0.858	0.712	0.048
Kurtosis	36.753	202.351	14.282	40.476	70.778	85.511	14.140
Skewness	4.168	11.660	0.387	1.298	5.796	5.847	0.538
J-B stat	29,713.4***	86,8012.6***	4,272.8***	34,395.8***	81,255.7***	105,867.6***	4,200.0***
Intraday Data							
Log-prices intraday							
N	3158	3171	3171	3157	3171	3171	
Mean	4.781	2.784	2.316	1.588	2.254	2.967	
Std	0.873	1.030	0.204	0.164	0.247	0.139	
Min	2.501	0.675	1.746	1.323	1.858	2.707	
25%	4.835	2.169	2.193	1.411	2.063	2.847	
50%	5.159	2.675	2.324	1.623	2.156	2.978	
75%	5.296	3.687	2.421	1.747	2.425	3.073	
Max	6.151	4.212	3.341	1.992	3.150	3.737	
Kurtosis	0.701	-1.071	1.308	-1.627	-0.383	0.573	
Skewness	-1.476	-0.477	0.258	-0.091	0.837	0.588	
J-B stat	1210.36***	271.69***	259.74***	352.14***	389.35***	225.17***	
Log-differences							
Mean	0.001	0.001	0.000	0.000	-0.000	−0.000	
Std	0.043	0.035	0.018	0.010	0.018	0.012	
Min	−0.693	−0.501	−0.293	−0.228	−0.121	−0.275	
25%	−0.008	−0.009	−0.005	−0.002	−0.006	−0.003	
50%	−0.000	−0.001	−0.000	0.000	−0.001	−0.000	
75%	0.008	0.008	0.005	0.002	0.006	0.003	
Max	0.754	0.842	0.282	0.250	0.395	0.179	
Kurtosis	106.281	142.306	68.638	277.062	76.591	120.167	
Skewness	1.084	3.993	1.856	2.681	3.637	−2.416	
J-B stat	1.481e6***	2.675e6***	622.093***	1.006e7***	779.316***	1.904e6***	

ber 2021. Regarding the analysis of comovements between the most mentioned stocks, 30-min intraday stock quotes data are employed. Finally, VECM is estimated using both

the intraday 30-min and daily news sentiment data.

Tab. 5: Volatility series

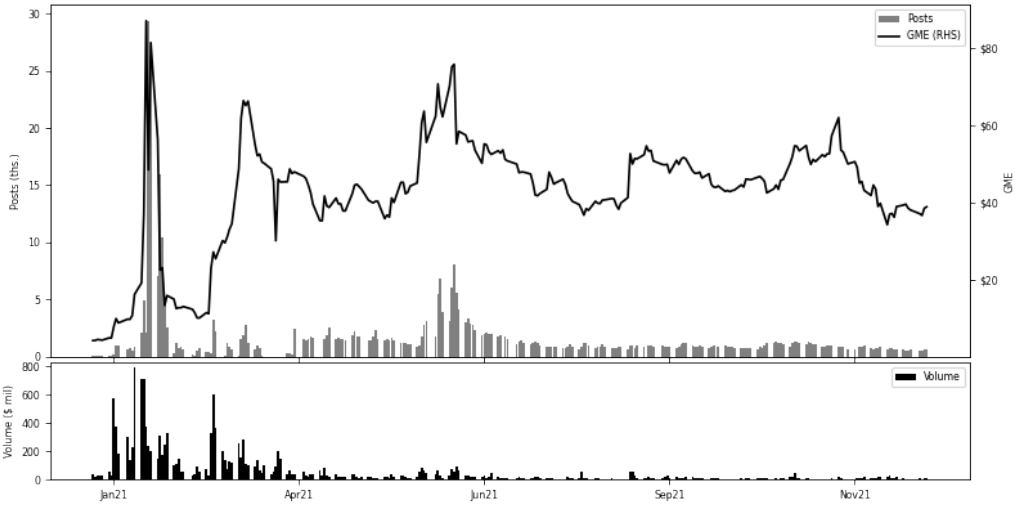
Descriptive statistics of the volatility series. Historical volatility is computed on a simple 10-day rolling basis. Implied volatility series reflecting options market proxying the speculative trading activity is downloaded from Bloomberg. Data covers the period from November 2020 until November 2021. J-B stat represents Jarque-Bera test statistic and *,** ,*** denote the 10%, 5% and 1% significance levels, respectively.

	GME		AMC		BB	
	Hist Vol	Implied Vol	Hist Vol	Implied Vol	Hist Vol	Implied Vol
N	252	252	252	252	252	251
Mean	158.60	160.76	158.62	175.47	83.91	95.63
Std	174.66	90.75	173.47	76.12	67.48	48.86
Min	27.48	68.22	26.66	94.17	22.38	52.63
25%	58.05	102.62	77.49	130.29	46.88	67.83
50%	88.81	130.15	114.21	157.98	66.26	80.27
75%	167.67	176.97	163.72	189.17	88.77	106.79
Max	999.03	538.34	1016.67	779.49	403.99	454.56
Kurtosis	9.01	3.45	14.25	18.43	11.91	16.63
Skewness	2.82	1.87	3.67	3.37	3.28	3.50
J-B stat	1144.05***	262.77***	2600.14***	3888.96***	1872.68***	3277.02***

4 Results

In the following section, we present the results of our analysis, divided into three parts. First, the results of textual and sentiment analysis are shown. Second, we discuss the wavelet analysis pertinent to the effect of Reddit sentiment, news sentiment and examination of the comovement. Lastly, following the results of the wavelet analysis, we review the outcomes of the vector error correction models.

Fig. 3: Activity
Bars represent the number of submissions in upper chart. Traded volume is included in the below chart. As indicated, more pronounced periods follow large price-volatility. The market activity peaked during January-May period.



4.1 Text analysis

We begin with purely semantic analysis. Fig. 3 shows how the activity on the subreddit corresponded to the periods when GME was most volatile and traded on the market. Using the whole sample of more than 438 thousand *r/wallstreetbets* unique submissions, word tokens, Porter's stems and lemmas are extracted. The overwhelmingly positive sentiment is evident at first sight as shown by Fig. 4, depicting the most frequent word tokens and bigrams. The prominence of *rocket, buy, hold, moon* or *let* word tokens indicates a large bullish sentiment. The broader context is presented by the bigram model extracting phrases such as *rocket rocket, let us, moon rocket, us go* or *hold line*. The bigram model is notably influenced by the translation of repeated emojis.

Moreover, highly frequent *short-squeeze, shares* or *short interest* terms manifest the relevancy of the short-selling phenomena in the trading frenzy, arguably one of its trig-

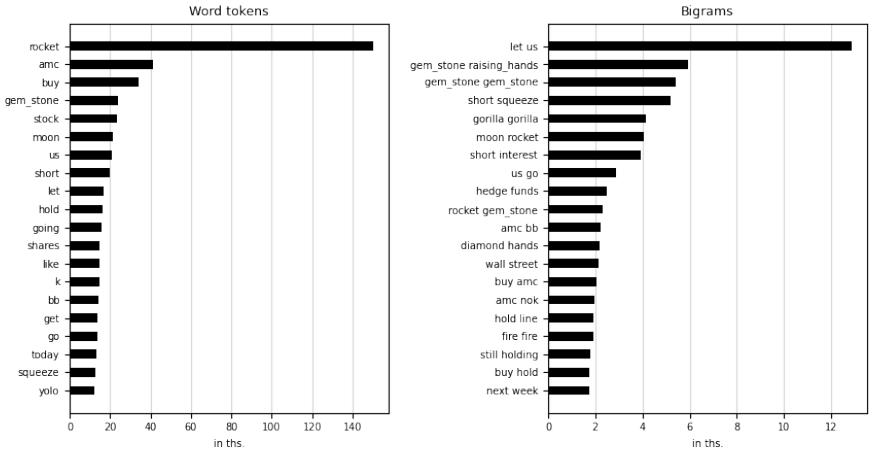
gering factors. The relevance of the selected sample of companies, shown by Fig. 5, demonstrates the large association users made between GameStop and other tickers.

4.1.1 Sentiment analysis

Moving beyond simple textual analysis, we assign sentiment scores to the scraped submissions. As discussed earlier, owing to the specific lexicon, predominantly used on the subreddit, we make necessary modifications to the lexicon to better reflect upon bearish and bullish sentiment.

Fig. 4: Word frequency

Left plot contains twenty most common occurring word tokens including translated emojis. The vertical axes contain the word token, respectively bigram, while horizontal axes denote the number they occurred in all of the submissions (in thousands). Regarding the bigram plot, the most frequent bigram *rocket rocket*, coming from the translation of consecutive corresponding emojis, is omitted for better visibility (it occurred more than 100 thous. times). Overall, the bigrams' frequency is heavily driven by the translated emojis



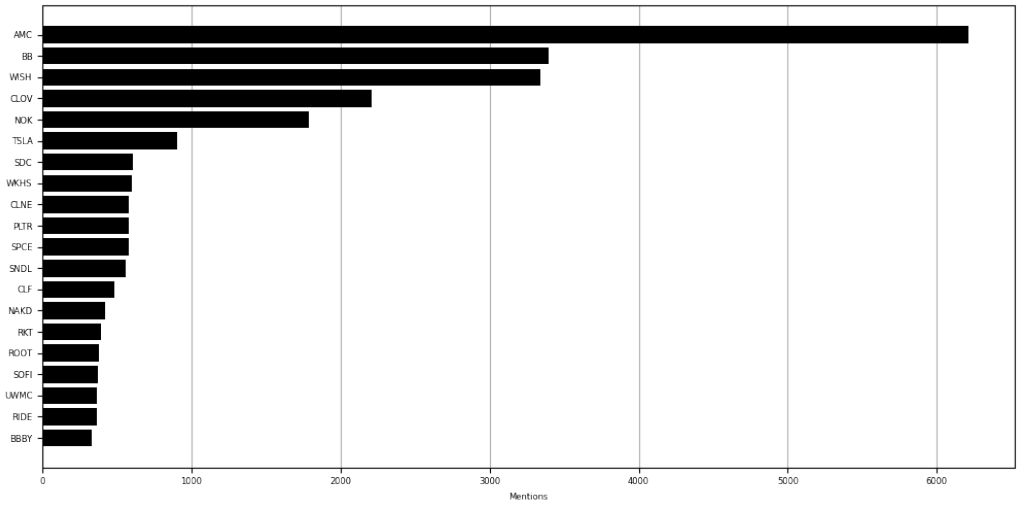
Interestingly, aggregating the scores intradaily, we detect large deviations and periods of strong negative sentiment, which almost disappear in aggregations of frequencies larger than six hours. The density plots of different sampling frequencies applied to the Reddit sentiment are included in Fig. A.2. This may suggest a fickle and reactive behaviour on the subreddit as the negative mood is quickly suppressed and drowned out by the trend of overwhelming positive sentiment. Furthermore, the frequency discerning ability of the wavelet analysis is highlighted.

Fig. 6, illustrating the intensity of news coverage and volatility, also identifies the two distinct periods in which sentiment surrounding GameStop peaked on the subreddit. We observe a large uptick in the news coverage activity that continues throughout the year. Moreover, there is a distinct difference between news and Twitter sentiment, which

is comparatively weaker than the period predating January 2021. Overall, all stocks experienced a surge in news mentions following the turn of the year, which coincided with large volatility.

Fig. 5: Most mentioned tickers

The most mentioned tickers alongside GME in all of the sampled subreddit submissions. The tickers are extracted using a simple convention that it must be preceded by a $ sign and contain up to six letters. Searching for ticker mentions we scrape both the title and the submissions' text body as opposed to only scraping titles, which is done to extract scores. Note that our sample choice is however driven mostly by January data, still the sampled stocks are heavily mentioned throughout the year as shown below.

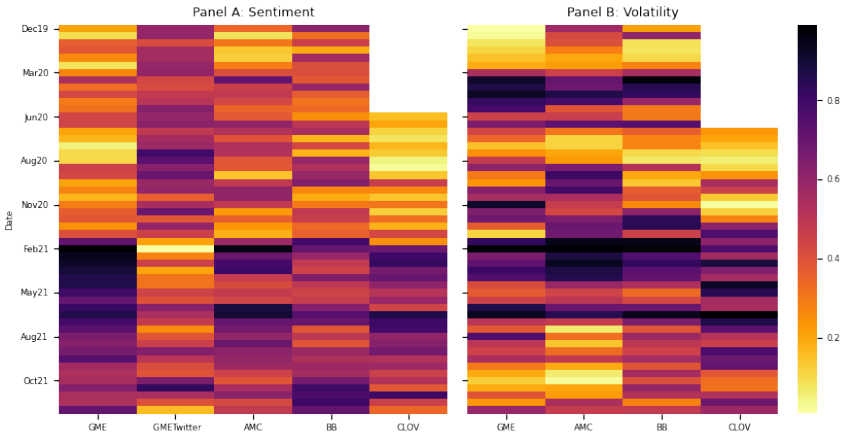


The comparison of the different sentiment datasets presented in our study is included in the Appendix (see Fig. A.2). Besides the large discrepancy in the respective dataset sizes, with social media understandably posting much higher numbers, we detect different patterns concerning the ratio of positive-to-negative submissions, respectively news. Reddit sentiment, as indicated by the sentiment analysis, is overwhelmingly positive, given that 50% of the posts are scored positive at all times. Whereas news sentiment is considerably more volatile and leaning to the extremes with prolonged periods when both positive and negative sentiment dominated. Furthermore, comparing the two social media, Twitter sentiment is much more varying compared to Reddit while being predominantly negative during the most active periods.

Nevertheless, all sentiment datasets pick up on the increased activity and heightened volatility starting from 2021, corresponding to the GameStop trading frenzy. Likewise, the news sentiment surrounding the other relevant stocks displays similar patterns, as shown by Fig. 6.

Fig. 6: Sentiment and volatility heatmaps

Following heatmaps portray the intensity of sentiment and volatility based on mean percentile ranks aggregated on a two weeks basis to achieve better visibility. The included colour scale denotes the percentile ranks. Panel A portrays the intensity of news coverage and twitter sentiment for GME, higher rank (closer to one) indicates greater quantity of news, respectively tweets, surrounding the stock. Analogically, Panel B describes volatility where higher volatility corresponds to darker colour. Missing values are left blank.



4.2 Wavelet Coherence

Let us now turn to the findings of the core tool of our analysis. Applying the wavelet framework to analyse the effects of Reddit investor sentiment during the GameStop retail trading frenzy, we discover interesting dynamics. Fig. 7 shows wavelet coherence plots between the sentiment and intraday 30-min returns, historical and implied volatility and, lastly, between the daily aggregated sentiment and GME returns.

First, we begin with a small guide on how to navigate the plots. The colour scale corresponds to the level of wavelet coherence, warmer (colder) colours signify higher (lower) coherence. The arrows represent the relative phase relationship between the two series, which can be indicative of causality. Their notation follows Torrence and Campo (1998) [81]. In-phase (anti-phase) relationship is denoted as $\uparrow$ ($\downarrow$), alternatively, it can be described as positive (negative) correlation. The phase difference is described by the horizontal plane of the arrows' notation, where a right-pointing $\rightarrow$ ($\leftarrow$) arrows suggest that the first (second) series leads the second (first) by 90°. For illustration, in Panel A $\nearrow$ implies that sentiment and GME returns are in-phase and the former leads the latter by 45°. In other words, sentiment positively influences GME. The horizontal axis corresponds to the time dimension while the vertical axis represents the frequency dimension, respectively the inverse of the period. For instance, the scales (periods) ranging from 2

up to 512 intervals in Panels A and B correspond to 1-256 trading hours. White contours designate time-scale areas where wavelet coherence is statistically significant at a 5% level, given by the Monte Carlo simulations procedure, as noted earlier. Finally, the cone of influence (COI) is given by white blurred areas wherein the edge effects can be effectively ignored.

Fig. 7: Wavelet analysis: Reddit Sentiment

Panel A, and B show wavelet analysis between the scraped Reddit sentiment and intraday GME returns and volatility (10-interval rolling standard deviation). Panel C and D use daily aggregated data and show wavelet coherence between $Sent$ and daily returns, and implied volatility. The scales on the y-axis correspond to either 30-min or 1-day intervals, i.e. in the upper (lower) panels 16-64 frequency bands correspond to 8-32 hours (trading days) or approximately 1-4 trading days. The $\nearrow$ ($\swarrow$) implies that the first (second) series leads the second (first) series and they are positively (negatively) related.

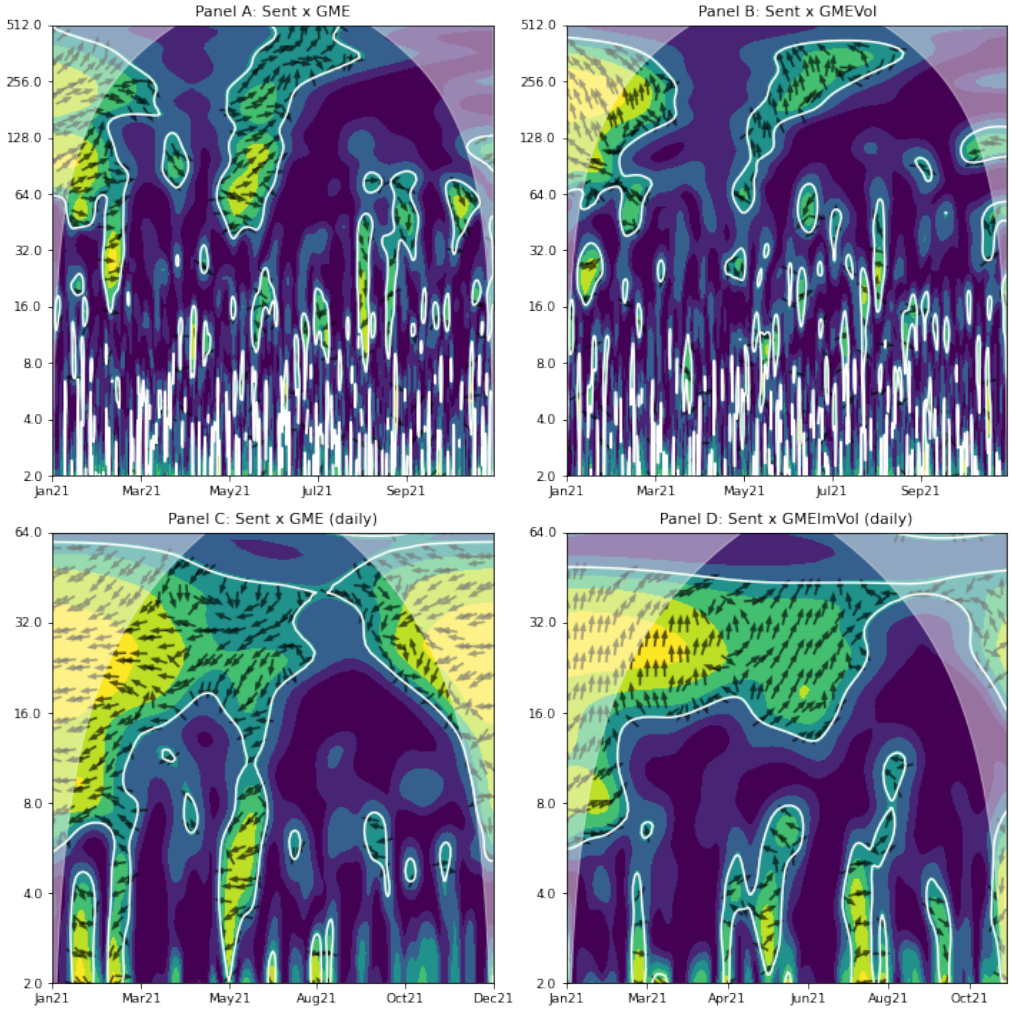


Fig. 7 Panel A and B show several short-lived areas of significant coherence at higher frequencies which can, however, occur by chance. More telling are the large areas of coher-

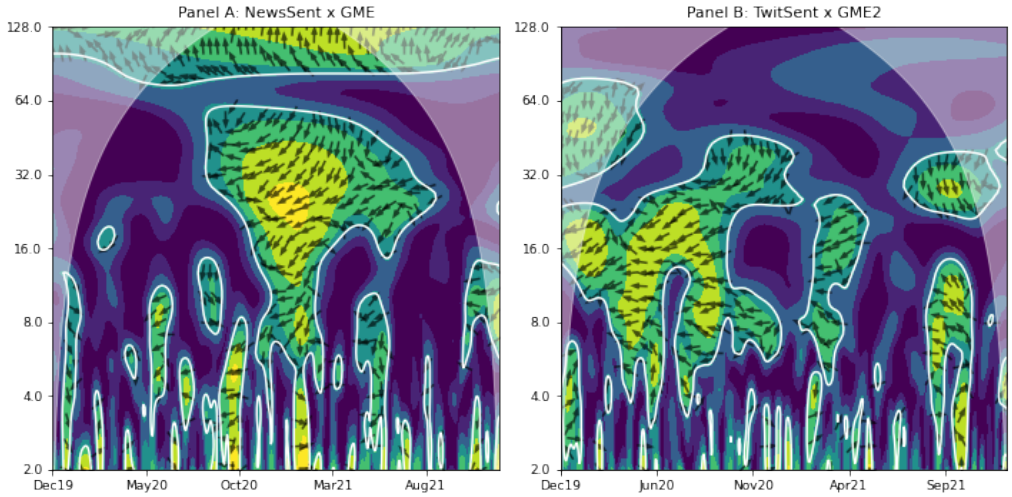
ence between the 8-128 scales corresponding to the most active periods on the subreddit. In Panel A, the two substantial periods occurring around February and May show two series largely in-phase with sentiment leading, suggesting that at 8-64 hours frequencies, corresponding to approximately 1-8 trading days, retail investor sentiment positively influenced GME returns. Beginning in August, there are also larger, albeit shorter-lived areas of significant dependence at high frequencies. Although, we observe a phase switch as sentiment and GME are interestingly anti-phase with no distinct phase difference. That is, larger sentiment had no effect on the returns, moreover, it was met with lower returns, in contrast to the previous periods of heightened activity on the subreddit.

Contrary to our expectation that sentiment is a better predictor of volatility rather than returns, we observe fewer and smaller time-scale areas of significant wavelet coherence, as shown by Panel B. Nonetheless, during the most active periods, we can likewise identify larger areas of significant dependence at 8-64 hour frequencies (16-128 scales). Sentiment and intraday volatility are largely in-phase with a small bias towards the former leading. Although, there is limited evidence for a switch in phase difference as volatility led sentiment in some periods, notably in June and towards the end of the year.

We also aggregate the Reddit sentiment daily (see Panel C). Expectedly, the high-frequency areas correspond to the 30-min intraday plot. At very low frequencies around 32-64 days, which are mostly excluded in Panel A, we detect large and significant coherence. However, the revealed dependence, as opposed to Panel A, signifies opposite phase difference and slightly anti-phase behaviour. The difference in the phase relationships depending on the sampling frequency may point to the fact that retail investors, while able to drive and influence short-run dynamics - in our case extremely short as we assess 30-min periods - the influence fades in the long-run as the phase turns. Still, we do not observe the different phase relationship examining volatility which is consistently positively related to the sentiment, as expected given the behavioural tendencies of individual investors characterised in the literature, see Odean (1999) or Barber and Odean (2008) [4, 14]. Thus, sustaining the argument that greater volatility spurs more discussion and feeds the growing sentiment on the subreddit. Further, the coherence between sentiment and implied volatility validates the highlighted role of options data, as argued by Fusari et al. (2020), Allen et al. (2021) or Baltussen et al. (2021) [70, 72, 73]. Coinciding with the volatility spikes of GME, extensive coherent time-scale regions on 2-8 wavelengths (days) are detected. Notably, at lower frequencies (>16 days), we observe consistent in-phase coherence which disappears after August despite the two series exhibiting dependence at higher frequencies.

Fig. 8: News and Twitter sentiment

Wavelet analysis between sentiment, news and twitter sentiment paring using daily data from December 2019 to December 2021. Vertical axis corresponds to period in days and horizontal axis corresponds to time. White contours denote 5% significance against red noise. The arrows represent relative phase difference. Cone of influence (COI) is delineated by white blurred area where edge effects take place.



News Sentiment

Provided that the news and Twitter datasets compared to the Reddit scraped sentiment extend further to December 2019, it additionally supplements unique findings. The figures are included in Fig. 8. Considering the nature of the news sentiment, as shown when comparing the sentiment series (A.2), we consider the news publication count. In the case of Twitter sentiment, the ratio of positive-to-negative tweets presents the best alternative to our Reddit scraped sentiment scores. News sentiment in Panel A similarly reveals large time-scale regions across all frequencies with significant coherence, especially from October 2020 onwards. Extending the sample to include the previous year shows that the dependence in the 8-64 frequency bands (days) occurred even before January 2021. Interestingly, it is shown, that news sentiment and GME are anti-phase with the latter leading, suggesting that rather than news releases containing fundamental information for GME, they were in reaction to the GME development in that period.

Panel B with Twitter sentiment allows for a more direct comparison to our scraped measure. Twitter sentiment and GME are shown to be highly dependent on investment horizons from one week up to a month, however, most of the dependence is found predating January 2021. While a strong coherence appears at the beginning of the year, it gets considerably weaker throughout the year 2021. Similarly to news sentiment, the phase arrows suggest that Twitter sentiment is led by GME moreover, anti-phase relation is

displayed. Considering the given results of the wavelet analysis, we argue that Reddit sentiment carries a distinct value compared to the news and even Twitter sentiment and that there is possibly a different relationship between the respective sentiments and the GME returns, at least suggesting that Reddit retail investors could have impacted the price development of GME.

4.3 Comovements

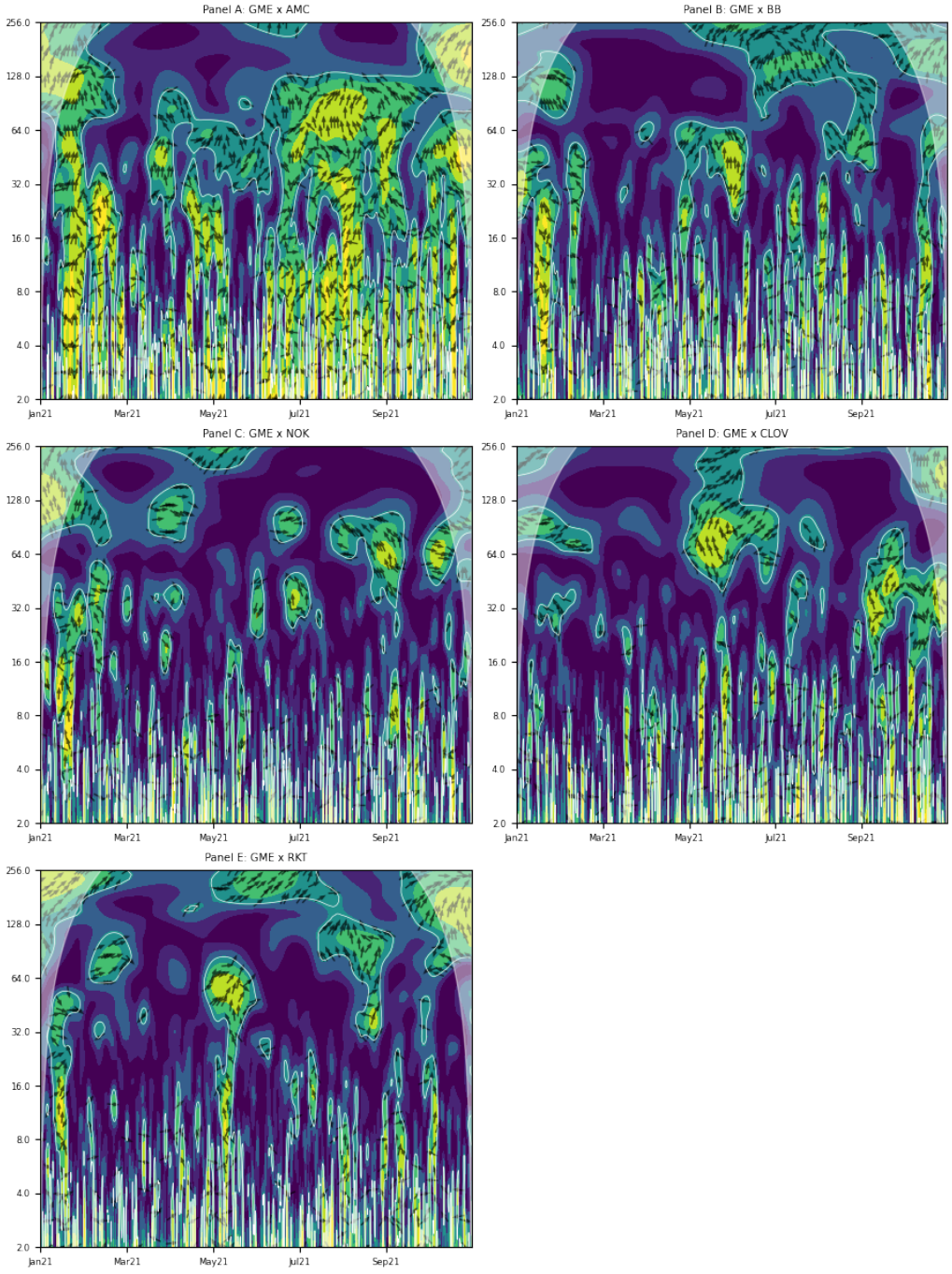
Finally, we employ the wavelets framework to analyse the contagion of the GameStop saga relevant to the selected sample of stocks discussed earlier. In addition to these most mentioned tickers surrounding the GME discussion on the subreddit, market-wide and short interest indices are used to assess the broader contagion. The results are reported in the respective Panels A-E in Fig. 9. AMC and GME exhibit by far the largest comovement, as demonstrated by the predominantly yellow and green coloured time-scale regions in Panel A, indicating large and consistent coherence. Both stocks show a very high degree of dependence even at higher frequencies, which is even more intense following June. As expected, considering the results of the textual analysis, both are in-phase. However, no distinct phase difference suggests that the possible spillovers went both ways, feeding off each other.

The other stocks do not display nearly the same level of comovement as AMC. Nevertheless, there are large significant time-scale regions where comovement occurs even at higher frequencies around 4-16 scales (2-8 hours). These time spans primarily coincide with the most active periods in the GME saga as well as the volatility spikes of the respective stocks. All series predominantly exhibit in-phase relationships. Moreover, there is no unequivocal evidence as to whether spillovers came from GME or the other way around, as indicated by the phase arrows. Therefore, apart from AMC, the contagion is largely limited to the periods of higher activity on the subreddit, during which these stocks experienced large price and volatility spikes. While the comovement at higher frequencies is relatively confined, the results suggest that retail investors' activity was closely associated with volatility and return spillovers.

Let us now turn to a broader context examining the market-wide and short interest indices comovements, see Fig. 10. GME and the short interest index do not exhibit larger significant comovement until the end of 2020. Already starting October 2020, we observe significant comovement which covers a larger frequency band about 2-16 days. Interestingly, during this period, at lower frequencies, the short index leads GME, while at higher frequencies phase difference turns opposite. Despite significant short-interest in GME, we are not able to identify a strong and consistent in- or anti-phase relationship,

Fig. 9: Comovements

Wavelet coherence of AMC, BB, NOK, CLOV and RKT with GME. Based on intraday 30-min data. Vertical axis runs from 1 to 128 hours, i.e. up to 16 trading days. Thin white contours denote 5% significance level against red noise. The arrows represent relative phase difference. Cone of influence (COI) is delineated by white blurred area where edge effects take place.

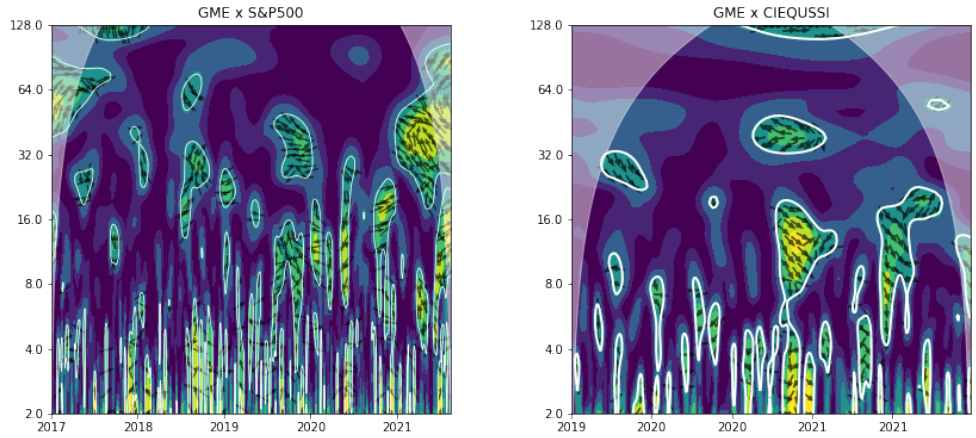


although, for a brief period in July, there is a strong anti-phase comovement. The differing phase difference and lack of phase relationship might lay evidence that GME returns were

driven, albeit for a short period, by retail investors' activity leading to deviation from the fundamentals.

Historically, there is understandably significant comovement between GME and S&P 500 index even at higher frequencies, although relatively short-lived. The increase in comovement corresponds to larger market declines, leading to a reduction in market complexity as multi-correlation between all stocks unilaterally increases given that common drivers arise, such as macroeconomic conditions. This is documented by Martina et al. (2011) [88] which is likewise mentioned by Vácha and Baruník (2012) [89] who assess the comovement of energy commodities. Following the sparse coherence periods in 2020, we observe a large white contoured time-scale region signalling in-phase dependence with GME leading. However, given the market volatility in that period, the historical pattern of comovement and the cone of influence, it cannot be reasonably concluded that this is attributable to GME saga spillovers.

Fig. 10: Broad market and short-interest index
Wavelet coherence plot of GME with broad market S&P 500 and Citi short interest equity index CIEQUSSI. Both series are sampled daily but the former extends back to 2017 while the latter only to 2019. Same description as in the previous plots applies.



In summary, we identify extensive time-scale regions where GME and other relevant stocks exhibit strong and significant in-phase comovement. These periods also coincide with the peaking activity on the subreddit. Furthermore, unsurprisingly, during these periods, the sampled stocks saw substantial volatility and their prices nearing peaks. GME and AMC, in particular, display extremely strong and consistent comovement bond throughout the sample period across all frequencies, which is a unique finding in our sample. While the phase differences do not bear conclusive evidence regarding the direction of the spillovers, using the short-interest index and the broad market index, we

suggest that the comovements between the retail targeted stocks are unique and not a result of the broad market development.

Following our supposition regarding sentiment being a better predictor of volatility rather than returns, we assess volatility spillovers in addition to returns, the figures are included in Appendix Fig. A.3. The results are mostly similar. We do not observe any substantial departure from the above findings. The coherence plots show abundant short-lived time-scale regions of significant comovement at higher frequencies which, however, could occur by chance or due to the broad market sentiment. The larger areas stretching over all relevant frequencies show that higher GME volatility is accorded with higher volatility of other stocks. Likewise, there is no conclusive evidence that spikes in GME volatility spilt over and directly caused spikes in the other discussed stocks.

4.4 Vector error correction model

We begin with a preliminary analysis of the data and the specification of the model. First, using the Augmented Dickey-Fuller (ADF) [90] and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) [91] tests, we check the order of integration, respectively unit root and stationarity. The results of the tests for the variables covered are included in Tab. 6. In the ADF test without a constant, the null hypothesis of unit root is rejected for all logarithmic price series, which is not the case when including a constant and a trend. Moreover, the sentiment series in any variation of the test appears to not have a unit root. On the other hand, KPSS tests reject stationarity for all series at a 1% significance level. Following the first differencing, ADF tests both with and without deterministic terms reject unit root for all variables. While ADF tests do not offer very convincing results, failing to reject unit root in some cases, KPSS tests still reject stationarity, indicating that the series might be at least close to a unit root. Moreover, after first differencing we are able to reject unit roots. Provided that, we proceed with the further specification of the model, following the procedure set out in Filip et al. (2019) [85].

The optimal lag structure (k) in Eq. (1) is determined by estimating VAR models with up to 12 lags. The choice of $k = 2$ is based on the Bayesian information criteria (Schwarz, 1978 [92]) to maintain parsimony. The other information criteria Akaike (AIC) (Akaike, 1974 [93]) and Final Prediction Error (FPE) (Ljung, 1998 [94]) preferred the maximum lag length of 12 or 6 in the case of Hannan-Quinn (HQ) (Hannan and Quinn, 1979 [95]). The choice of optimal lag is robust to both exclusion and inclusion of deterministic trends. Moreover, given the high frequency of our data, note that, the lag order of two represents very short-term dynamics. Next, following the Johansen (1991) [86] procedure the number of cointegration relations is tested. Tab. 7 presents the results using both the *trace* and

Tab. 6: Stationarity tests
The p-values of Augmented Dickey-Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) are enclosed. All three variations with respect to deterministic terms of the former are presented, the optimal lag selection is based on the Bayesian information criterion (BIC). Results for both original data and after first differencing are included.

	ADF (no drift)	ADF (drift)	ADF (trend)	KPSS
Intraday data				
Sent	<0.01	<0.01	<0.01	<0.01
GME	0.5560	<0.01	<0.01	<0.01
AMC	0.3608	<0.01	0.0235	<0.01
BB	0.7883	<0.01	<0.01	<0.01
NOK	0.4023	0.0723	0.0389	<0.01
CLOV	0.3164	<0.01	<0.01	<0.01
RKT	0.7180	0.0165	<0.01	<0.01
Daily data				
GME	0.6763	<0.01	<0.01	<0.01
news	0.4018	<0.01	<0.01	<0.01
Differenced			.	
GME	<0.01	<0.01	<0.01	.
Sent	<0.01	<0.01	<0.01	.
AMC	<0.01	<0.01	<0.01	.
BB	<0.01	<0.01	<0.01	.
NOK	<0.01	<0.01	<0.01	.
CLOV	<0.01	<0.01	<0.01	.
RKT	<0.01	<0.01	<0.01	.
GME	<0.01	<0.01	<0.01	.
news	<0.01	<0.01	<0.01	.

maximum eigenvalue statistics, based on which two cointegrating relations are found. Both test statistics are unable to reject the existence of more than two cointegrating vectors at a 1% significance level, thus, we conclude that there are two cointegrating relationships among the variables, i.e. $r = 2$. The suggested two long-run equilibria would be consistent with the existence of a broad market-driven equilibrium among the variables and the relationship between sentiment and the stocks driven by retail investors. Note that, both constant and a trend are excluded from the error correction terms, in which case unit roots are rejected for both error correction terms. That is, we arrive at our final specification of the VECM(1) with the unrestricted constant variation as per Henry and Juselius (2001) [84], including GME, *Sent*, AMC, BB, NOK, CLOV and RKT, and two cointegrating vectors.

The results of the system of equations are summarised in Tab. 8. Given the high frequency of the data, the one lag in the respective equations represents very short-term dynamics. We can observe that on a very high-frequency basis *Sent* is a significant determinant for most of the stocks, however, with a varying sign. The effect of sentiment scores is shown to be positive for GME, AMC, NOK and RKT while negative for others.

Tab. 7: Johansen cointegration tests
The tests' specification contains one lag as suggested by Bayesian information criterion (BIC) and a constant outside the cointegrating relation. Both trace (LR_{trace}) and maximum eigenvalue statistics (LR_{max}) are supplied including the critical values for 1% and 5% significance levels.

Rank	LR_{trace}	$CV_{5\%}$	$CV_{1\%}$	LR_{max}	$CV_{5\%}$	$CV_{1\%}$
r ≤ 6	3.305	8.180	11.6500	3.305	8.180	11.650
r $\leq$5	8.916	17.950	23.520	5.611	14.900	19.190
r $\leq$4	22.907	31.520	37.220	13.991	21.070	25.750
r ≤ 3	46.068	48.280	55.430	23.161	27.140	32.140
r $\leq$2	78.645	70.600	78.870	32.578	33.320	38.780
r $\leq$1	137.048	90.390	104.200	58.403	39.430	44.590
r=0	348.795	124.250	136.060	211.747	44.910	51.300

Regarding the error correction terms, at least one of the two is negative and significant for GME, *Sent* and BB, indicating adherence to some long-run equilibria. In other cases, we lack either the requirement that the error-correction term is negative, i.e. adjustment for previous deviations from the equilibrium, or statistical significance of the term, i.e. zero convergence to equilibria.

Tab. 8: VECM results
Matrix of coefficients of the respective VECM equations, t-stats supplied in parentheses. Note that *Sent* score is divided by 100 for better scaling. *ect* denote error correction terms, the numbers in parenthesis in the name of the variable indicate the lag order.

	GME	*Sent*	AMC	BB	NOK	CLOV	RKT
ect_1	−0.0028 (−0.7398)	−0.0028 (9.1848)	0.0041 (1.3559)	−0.0042 (−3.0658)	0.0005 (0.8775)	−0.0003 (−0.1676)	0.0011 (1.1337)
ect_2	−0.0250 (−1.9279)	−0.2046 (−13.7934)	0.0013 (0.1276)	0.0061 (1.2952)	−0.0023 (−1.1484)	0.0139 (2.6430)	0.0048 (1.4046)
constant	−0.0290 (−2.3190)	−0.1875 (−13.1255)	0.0066 (0.6633)	0.0029 (0.6524)	−0.0019 (−0.9690)	0.0148 (2.9085)	0.0062 (1.9052)
$GME(-1)$	−0.0082 (−0.2857)	0.0057 (0.1732)	0.0106 (0.4657)	0.0202 (1.9604)	0.0069 (1.5256)	0.0081 (0.6961)	−0.0026 (−0.3502)
$Sent(-1)$	0.0570 (3.4350)	−0.3342 (−17.6102)	0.0684 (5.1863)	−0.0109 (−1.8274)	0.0021 (0.8087)	−0.0231 (−3.4264)	0.0013 (0.2892)
$AMC(-1)$	−0.0966 (−2.6114)	−0.0200 (−0.4729)	−0.1236 (−4.2046)	−0.0155 (−1.1613)	−0.0299 (−5.1338)	−0.0006 (−0.0404)	−0.0157 (−1.6260)
$BB(-1)$	0.2071 (3.1694)	−0.1558 (−2.0855)	0.1470 (2.8315)	0.0243 (1.0339)	0.0457 (4.4346)	−0.0133 (−0.4998)	0.0370 (2.1639)
$NOK(-1)$	0.1526 (1.0312)	0.3010 (1.7791)	−0.2089 (−1.7760)	0.0700 (1.3141)	−0.0838 (−3.5934)	−0.0159 (−0.2651)	−0.0005 (−0.0118)
$CLOV(-1)$	0.0923 (1.7776)	0.1401 (2.3588)	0.0296 (0.7170)	0.0210 (1.1230)	0.0210 (2.5665)	0.0640 (3.0318)	0.0088 (0.6501)
$RKT(-1)$	−0.0023 (−0.0293)	−0.0465 (−0.5162)	−0.0672 (−1.0736)	−0.0092 (−0.3260)	−0.0038 (−0.3088)	−0.0399 (−1.2452)	0.0299 (1.4523)

To test the respective exogeneity of the variables we test the restrictions on the loading coefficients (α), i.e. the speed of adjustment, using the likelihood ratio tests, the results can be found in Tab. 9. Further to examine short-term dynamics we utilise the concept of Granger causality in the VAR representation of our VEC model. Tab. 10 contains the p-values of the respective (conditional) Granger causality tests, where the null hypothesis is stated as *the variable in the row does not Granger cause the variable in the column.* The exogeneity tests and short-run dynamics offer interesting results. There is strong evidence that *Sent* Granger causes GME, AMC and CLOV while being Granger caused by BB,

NOK and CLOV. Surprisingly, the GME position in the system is relatively weak as it only shows bi-directional causality with BB in addition to being affected by *Sent*, AMC and CLOV. Thus, it is shown that GME development alone does not explain *Sent* however, the opposite direction applies. Meanwhile, AMC which exhibited strongest comovement with GME in the wavelet analysis, appears to affect GME and NOK while being affected by *Sent*, BB and NOK. The relatively weakest link in the system is RKT which seems to have no effect on other variables while only being affected by BB. Regarding the exogeneity tests we are not able to reject the null restriction on the loading coefficients at a 5% level only for AMC, NOK and RKT, which are then considered weakly exogenous. Thus, it is shown that GME, *Sent*, BB and CLOV are largely driving the short-run dynamics in the system.

Tab. 9: ECT terms

The p-values of tests of null restrictions on cointegration parameters are supplied, including loading (α) and long-term relationship (β) coefficients. The latter two coefficients are standardised to GME and *Sent*. The cointegrating vectors indicate the long-term relationships between the variables. Note that the standardised cointegrating vectors do not include GME and *Sent* in the respective standardised vectors.

	GME	*Sent*	AMC	BB	NOK	CLOV	RKT
α p-value	0.0275	<0.01	0.314	0.0415	0.5171	0.0129	0.092
β_{GME} p-value	.	.	<0.01	<0.01	0.078	0.226	<0.01
β_{Sent} p-value	.	.	0.741	0.432	0.345	0.607	0.026

Cointegrating vectors

$$GME = 0.556 \cdot AMC - 1.595 \cdot BB + 0.654 \cdot NOK - 0.094 \cdot CLOV + 2.440 \cdot RKT$$

$$Sent = 0.052 \cdot AMC - 0.197 \cdot BB + 0.049 \cdot NOK + 0.287 \cdot CLOV + 0.261 \cdot RKT$$

Now, moving to the long-term relationship, the cointegrating relations are examined. We obtain two cointegrating vectors normalised to GME and *Sent* but note that neither of them figures in the respective cointegrating vector. Given the log-form, betas in the cointegrating relationship normalised to GME can be interpreted as long-term elasticities. Regarding the directional influence in the long-run equilibria, we observe that apart from BB and CLOV all other stocks are positively related. Note that, however, the bivariate wavelet analysis for all GME pairs has shown us an in-phase relationship in the corresponding lower frequency bands. Notably, AMC and RKT are negatively related in the short-run (see Tab. 8) while having a positive sign in the cointegrating relation. The second cointegrating vector normalised to *Sent* shows that in the long run all the most discussed stocks apart from BB are positively related to the sentiment on the subreddit. For parsimony, we also attempted to run the VECM with one cointegrating relationship which results, however, we do not enclose here. In such specification besides yielding similar results *Sent* is shown to be negative and statistically significant in the linear com-

bination normalised to GME, i.e. the cointegrating relationship, which is consistent with the in-phase relationship found in the wavelet framework, using daily data.

Tab. 10: Granger causality
p-values of Granger causality tests for short-run dynamics for Reddit sentiment and the sampled stocks. The tested null hypothesis is represented as *the variable in the row does not Granger cause the variable in the column* Given that our specification includes only one lag, these F-tests boil down to t-tests in the respective equations.

	GME	*Sent*	AMC	BB	NOK	CLOV	RKT
GME		0.8625	0.6415	0.0499	0.1271	0.4864	0.7262
Sent	<0.01		<0.01	0.0676	0.4187	<0.01	0.7725
AMC	<0.01	0.6363		0.2455	<0.01	0.9677	0.1039
BB	<0.01	0.0370	<0.01		<0.01	0.6172	0.0305
NOK	0.3025	0.0752	0.0757	0.1888		0.7909	0.9906
CLOV	0.0755	0.0183	0.4733	0.2614	0.0103		0.5156
RKT	0.9766	0.6057	0.2830	0.7444	0.7575	0.2131	

Finally, given the high frequency of the data, we also enclose the results of the same VEC model using daily aggregated data on Reddit sentiment, specifically the Granger causality tests indicative of short-run dynamics, see Tab. A.1 in Appendix. The other results are largely similar to the ones presented above concerning both the VAR coefficients and long-term cointegrating vectors. As opposed to GME possessing a relatively weak position intraday, using lower frequency data, it appears to be a substantial driving force for both *Sent* and other mentioned stocks. Therefore, further underlining the difference in the relationship based on different investment horizons as emphasised by the wavelets framework.

4.4.1 News Sentiment

Lastly, following our interest in comparing the news and social media sentiment, we also obtain results from the VECM using daily news sentiment data. Our comparison is rendered comparably weaker due to the different sampling frequencies. Moreover, we only estimate a bivariate model including news sentiment and GME for more befitting comparison to the bivariate wavelet analysis. Using the whole sample from December 2019 until December 2021, however, we fail to identify cointegration in the system, presumably due to the large structural break at the beginning of 2021. Notice that our Reddit scraped sentiment dates back to only January 2021. Therefore, the dataset is truncated to only include the year 2021. Following which, the Johansen test [86], both based on *trace* and *maximum eigenvalue* statistics, finds evidence for one cointegrating vector.

The preliminary analysis and model specification follows analogically all the steps presented above, therefore, let us skip to the discussion of the results. The estimated VECM(4) includes one cointegration and a constant outside of the cointegration space, the results are summarised in Tab. 12. The error-correction term is significant and

Tab. 11: Johansen cointegration tests
The tests' specification contains four lags as suggested by Bayesian information criterion (BIC) and no deterministic terms in the cointegrating relationship. Both trace (LR_{trace}) and maximum eigenvalue statistics (LR_{max}) are supplied, including the critical values for 1% and 5% significance levels.

Rank	LR_{trace}	$CV_{5\%}$	$CV_{1\%}$	LR_{max}	$CV_{5\%}$	$CV_{1\%}$
r $\leq$1	2.75	8.18	11.65	2.892	9.240	12.970
r=0	35.09	17.95	23.52	32.792	15.670	20.200

negative only for GME, suggesting that the news count does not converge to the equilibria as it is weakly exogenous. The short-run dynamics given by the Granger causality are manifested to run only in one direction as GME Granger causes news, but not the other way around, see Tab. 13. Looking more closely at the VECM results, GME appears to positively influence news, or in other words, better performance of GME gains more traction in the news. Meanwhile, the long-run relationship, given by the cointegrating vector, shows that higher GME returns unexpectedly correspond to lower news coverage. The effect is, moreover, highly statistically significant.

Tab. 12: News VECM results
Matrix of coefficients of the respective VECM equations, t-stats supplied in parentheses. ect represents error correction terms, the numbers in parenthesis in the name of the variable indicate the lag order. The beta coefficient standardised to news, indicative of long-term relationship, is also included.

	GME	news
ect_1	−0.1025 (−5.6284)	−0.0911 (−0.985)
$constant$	0.6717 (5.671)	0.579 (0.963)
$GME(-1)$	0.0004 (0.006)	0.948 (3.0421)
$news(-1)$	0.0226 (1.7473)	−0.5118 (−7.795)
$GME(-2)$	0.1831 (3.0123)	1.1119 (3.6032)
$news(-2)$	0.0124 (0.8922)	−0.519 (−7.3826)
$GME(-3)$	0.2534 (4.058)	0.7863 (2.4807)
$news(-3)$	0.0064 (0.4682)	−0.3283 (−4.7056)
$GME(-4)$	−0.2333 (−3.636)	0.3168 (0.9726)
$news(-4)$	0.0053 (0.4288)	−0.3198 (−5.0746)
EC coefficients	value	p-value
β_{news}	−0.2904	<0.01

To conclude this section, we briefly comment on the similitudes and differences between the wavelet analysis and VEC models. Using the high-frequency data, both methods come to an agreement that retail investor sentiment is positively related to GME, moreover, that the sentiment appears to be driving the short-run dynamics. On the other hand, in contrast to the in-phase relationship in the wavelet framework when using intraday data

in frequency bands up to a week, the long-run estimated relationship between sentiment and returns is indicated as negative in VECM. This result is, however, consistent with the daily aggregated wavelet coherence in which we observe similar dependence as for news sentiment. The bivariate wavelet analysis indication of the strong role of AMC and BB is further confirmed in the multivariate VECM setting. VECM moreover expands on the detected strong in-phase wavelet coherence between GME and other most mentioned stocks, by testing short-run Granger causality which suggests a rather weak role of GME in the system. Nonetheless, the strong dynamics between the stocks are consistent with the strong wavelet coherence. While we focused on high-frequency areas using the wavelet analysis, VECM allowed us to detect long-run equilibria to which the variables relatively strongly adhere. Lastly, regarding news sentiment, while the phase difference is validated by the VECM suggesting that short-run dynamics point from GME to news, their implied direction is opposite. Although, in the long run, news are indicated to be negatively related to GME returns, laying further evidence for the anti-phase wavelet coherence.

Tab. 13: Granger causality tests
p-values and test statistics for Granger causality tests between news sentiment and GME are supplied.

Null hypothesis	χ^2 stat	p-value
NEWS does not Granger cause GME	3.155	0.532
GME does not Granger cause NEWS	34.264	<0.01

5 Discussion

Given the presented results, let us continue with a summary and further relate to the literature presented at the beginning to discern the contribution of our work. First, unlike other literature following the GameStop saga through the lens of individual investor sentiment, which employed various data from news and tweets count to options data, see Umar et al. (2021a, 2021b) or Fusari et al., (2020) [67, 68, 70], we obtain an endogenous measure of the sentiment using scraped data from social media. In particular, submission data from Reddit where the retail investors aggregated.

The semantic analysis of the obtained textual data offers intriguing results showing the overwhelmingly positive sentiment throughout the sample period, consistent with the results of Long et al. (2021) [69]. The textual and subsequent sentiment analysis allowed us to aptly capture the retail investor mood surrounding GameStop and other tickers, additionally identified as targeted by the Reddit investors. Provided evidence shows that the sentiment build-up and activity on the subreddit coincided with periods of largest volatility of the respective stocks. Nonetheless, while sentiment was comparatively muted, barring these periods, it was still unyieldingly positive. In comparison to the news sentiment, it is proven to be distinctively different, both from the perspective of volume and assessment, ascertaining that our measure of sentiment offers unique optics and information compared to the traditional source of sentiment. Furthermore, rather interestingly, it also somewhat differs from the mood on Twitter. In contrast to the traditional sources of investor sentiment, such as trade imbalances, news coverage, or even google trends, we argue that our measure captures, besides an authentic link between the market development and investor mood intraday, also a powerful endogenous measure, given the fact of assessing real-time discussion. Specifically aimed at individual investors, the sentiment derived from the vast volume of textual data provides deeper insight into the behavioural tendencies of retail investors, who stand in the spotlight of our work. Accordingly, to the representation of noise traders, the Reddit sentiment is relatively noisy both owing to the fickle nature of retail investors and high-frequency sampling. Despite that, it represents a pertinent and significant connection to market development.

Binding together the activity on the subreddit and the market in the bivariate wavelet analysis, large and significant dependence between sentiment and GME is detected during the most volatile periods, with the latter leading. Contrary to Atkins et al. (2018) [15] and our expectations, we find the connection stronger for the return series rather than volatility, on an intraday basis. The dependence between $Sent$ and GME and the periods at which the coherence occurs suggest that individual investors' attention is heavily skewed towards higher price volatility, extreme one-day returns and other attention-grabbing

biases as documented by Barber and Odean (2008) [14], Andrade, Chang, Seasholes (2008) [50], or Brandt et al. (2010) [51], among others. The wavelet coherence and phase difference between sentiment and GME are, moreover, is exhibited highly dependent on the investment horizon. While retail investors are capable of driving short-run dynamics for a brief period, both wavelet analysis and VECM cast doubt over stronger long-run dependence. That is, even though retail investor sentiment is demonstrated to have an effect, it is only so for brief periods at a time. Nevertheless, note that the large price deviations last markedly longer, given that, we do not observe complete and immediate reversal. That is largely consistent with the *noise trader theory* which describes the long-run deviation from fundamentals resulting from limited arbitrage and individual investors trading based on pseudo-signals, see Black (1996) and De Long et al. (1989, 1990a, 1990b) [12, 26, 27, 28]. The limits to the arbitrage are apparent given the short-sale restrictions and costs documented at the beginning of 2021 as evidenced by Allen et al. (2021), Pedersen (2021) or Umar et al. (2021) [72, 96, 66]. Related to the susceptibility to pseudo-signals, bivariate wavelet analyses suggest that the activity on the subreddit accords to the market activity as investors were lured by the volatility and captivated by the visions of quick profits, mostly confirming the behavioural tendencies described by Odean (1999), Barber and Odean (2000, 2001, 2002) or Kumar and Lee (2006) [4, 42, 43, 5, 6].

These behavioural proclivities, such as overconfidence, self-attribution bias or illusion of control, are arguably exacerbated by social media where these biases are reinforced by a large congregation of individual investors who then, essentially, happen to trade in unison. The largely self-perpetuating sentiment in the subreddit discussions vastly promotes the inclination of retail investors towards similar behaviour, similarly to the evidence procured by Jackson (2003) or Feng and Seasholes (2004) [45, 46]. Much like the theoretical *noise traders* act on the noise under the conviction that it represents fundamental news, the continuously repeating information, multiplying the sentiment on the subreddit, on its own, constitutes a major signal interpreted by the retail investors. This *echo chamber* effect, most likely present in our case, is described by Jiao et al. (2020) [10].

On the difference between news and social media sentiment, the wavelet analysis shows that the relationships with GME and other stocks are conditional on the investment horizon. While the high-frequency dataset indicates that sentiment and GME are in-phase in the short-run with a possible phase switch at lower frequencies, news sentiment and GME exhibit anti-phase behaviour in 8-32 days bands and in-phase relationship at very low frequencies (>80 days), meanwhile, VECM indicates the opposite. The observed significant wavelet coherence in some periods may suggest that Reddit sentiment affects GME, whereas news sentiment appears to be largely driven by the GME development. Hence, this suggests that news sentiment did not necessarily reflect the measure of fun-

damental news, which would have presumably been the key component for the price discovery. Rather, it is implied that news picked up on the market activity or that GME deviated from the fundamentals, the same is apparent from the VECM results. Ultimately, while both sentiments are shown positively related to GME in the short-run, however, with bi-directional Granger causality between *Sent* and GME as opposed to short-run dynamics only running from GME to news, the long-run relationship appears to be different. Interestingly, we find that both news and *Sent* are negatively related to GMe in the long run., as indicated by the cointegrating vectors. It might be that, while retail investor sentiment is proved to significantly influence the short-run dynamics, it generates substantial noise in the long run, leading to the effect dissipating. Whereas the negative relation between news and GME is relatively warring, it may indicate that retail investor activity has led to the deviation from the fundamentals. Notwithstanding, the differing short- and long-run dynamics would be consistent with the stylised fact that retail trades tend to push prices in their direction which is, however, followed by subsequent reversals over long horizons as shown by, for example, Andrade et al. (2008) or Barber et al. (2009) [50, 7]. However, admittedly, given the different sampling frequencies, terms *short-run* and *long-run* are slightly misrepresentative. Nonetheless, following Gidofalvi and Elkan (2001) and Antweiler and Frank (2004), Tetlock (2007), Schumaker and Chen et al. (2009), Chen et al. (2014) or Farrell et al. (2022) ([57, 8, 97, 60, 9, 11], we find a significant relationship between stocks targeted by retail investors and social media, and news sentiment. However, we cannot rightly conclude on the vastly different effects of social media and news sentiment on returns and volatility as illustrated by Jiao et al. (2020) or Atkins et al. (2020) [10, 15].

Regarding the spillovers from GME to other most mentioned stocks alongside it, there are large and significant comovements during the most active periods. Most notably, as presupposed by the textual analysis, AMC and GME show extreme levels of consistent comovement throughout the sample periods unmatched by the other stocks. The multivariate VECM setting further implies the existence of two cointegrating relations, presumably related, first, to the broad market and second, to the link between sentiment and other targeted stocks. The Granger causality tests indicate that *Sent* and AMC possess a relatively strong role in driving the short-run dynamics, whereas GME is shown to be largely determined by other forces including retail investor sentiment on a fairly short 30-min intraday basis. Essentially, the marked role in driving short-term dynamics, the strongest dependence occurring intermittently during periods characterised by large volatility and the differing long-run relationship between sentiment and GME, all would suggest that the effect of retail investor sentiment lacks breadth. Both for consistent long-run impact and the ability to target a broader set of stocks, given the limited con-

tagion primarily confined to only certain periods. Regardless, we confirm that the link between individual investor sentiment and GME extended to the relevant stocks around the retail investors rallied the most. While the connection between news and social media sentiment, per se, is not novel - given the extensive above-mentioned literature - more importantly, we emphasise the distinct value of social media sentiment in connection to the selection of stocks which were targeted by retail investors. In particular, retail investors aggregated on Reddit, which has shown to reliably predict stock returns before, as argued by Bradley et al. (2021) [65]. Despite the apparent behavioural tendencies in the subreddit, it is demonstrated that retail investors can substantially impact pricing and possibly create short-term imbalances. Albeit the effect is relatively short-lived and different from the possible estimated long-run relationship, the imbalances can be fairly sustained.

Limitations

Before moving to the conclusion, a concise discussion regarding the limitations of our study is provided. At the centre of our work stands the high-frequency data on the scraped sentiment and sampled stocks, while the news sentiment is only obtained on a daily frequency. Hence, pertinent to our interest in comparing them, some of the value is lost. Moreover, while we were able to tweak the scraped sentiment, to aptly reflect the specific nature of the subreddit discussion, we do not possess the same capabilities regarding the news sentiment data. The difference in methodology could further contribute to the diversion between the two. Further, our sample period in some cases is relatively short, especially when considering daily data limiting our inference, which, in particular, limits the wavelet analysis for which the high-frequency data are preferred. Further, using the scraped sentiment, although we are able to appropriately the mood and behavioural tendencies of individual investors, it can serve only as a proxy of their real market behaviour. That is, compared to the data used in the extensive literature attempting to describe retail investors' behaviour, including retail trade imbalances, brokerage account data, etc., see Kumar and Lee (2006), Kaniel et al. (2008) or Barber et al. (2009) [6, 48, 7], we are not quite able to make that direct connection relying on a proxy. Lastly, the results are documented predominantly for the stocks most occurring in the subreddit discussions. However, it is not of our interest to comment on the impact of retail investors on the broad market. Instead, we intentionally limit our scope to the specific stocks targeted by the retail investors, which are largely characterised by the dominant features appealing to the individual investors, as described by extensive literature.

6 Conclusion

In our study, we have focused on the role of individual investors in the financial markets, particularly, the impact of the retail investor sentiment. With this aim, we specifically target the events surrounding GameStop and the activity of retail investors on the social media platform Reddit from the beginning of the year 2021, principally owing to the unique features of the retail trading frenzy. Through the optics of individual investor sentiment, we have attempted to detect and eventually describe the connection between sentiment and the stock market. Central to the thesis is the endogenous measure of social media sentiment obtained from the interactions of retail investors in the community on the Reddit social media called subreddit. In addition to the traditional time-series methods, we employ the novel approach of wavelet analysis, allowing us to parse through both time and frequency space, examining the dependency of sentiment and market risk or returns, as well as the comovement of the stocks which were mostly targeted by the retail investors. Motivated by the findings of the wavelet analysis, a traditional time-series domain tool, in the form of a vector error correction model, is employed to assess and further distinguish between the short- and long-run relationships between the variables. The contribution of the thesis, as laid out in the beginning, is essentially threefold, which we shortly summarise hereafter.

First and foremost, at the centre of the thesis, stand the retail investors and the noise trader theory. The GameStop saga was exceptionally unique in regards to the limited arbitrage conveyed by the short interest and short sale constraints, in addition to the coordinated efforts of individual investors, which were continuously captured on the subreddit. The obtained textual data and consequently extracted sentiment reflect an endogenous and precise gauge of the retail investor mood, provided that the sentiment scores stem from analysing thousands of posts. One of our initial findings confirms that the intensity of the discussion on the subreddit coincided with the market activity. Respectively, evidence provided shows that increased volatility accorded with heightened activity and mood on the subreddit. Moreover, as opposed to news and Twitter sentiment, Reddit retail investors stayed predominantly positive throughout the sample period. The outcomes of the textual and sentiment analysis further exhibit the extensive noise in the subreddit discussion surrounding GameStop.

The conducted bivariate wavelet analyses reveal the distinctive behaviour of the individual investors, which is largely consistent with the previously mentioned literature. In particular, the appeal of highly volatile, extreme one-day returns, and news-mentioned stocks seem to captivate retail investors. We show that, on an intraday basis, the sentiment carries valuable information concerning the most mentioned stocks' returns and

volatility. Specifically, strong coherence between the variables across several frequency bands is detected. Most interestingly, sentiment and returns exhibit extensive dependence even at high frequencies from one day up to a week. Additionally, we find a significant, albeit very short-term, the impact of retail investor sentiment. While there are bi-directional relationships between the sentiment and other stocks, it possesses a comparatively strong role in the system. Concurrently, GME is predominantly the impulse variable, suggesting that the spillovers were not coming directly from GME, per se, but rather from the sentiment surrounding it, as retail investors targeted other relevant stocks. Anyhow, at lower frequencies, GME's role is much more pronounced as it largely contributes to driving the short-run dynamics, whereas sentiment is determined by the market activity, albeit still possessing significant predictive power. Considering the laid out results, we show that retail investor sentiment carries valuable information about predicting market returns and volatility. However, notably, retail investors lack breadth as wavelet analysis indicates only limited time spans in which coherence is highly potent, further confirmed by VECM. Moreover, the effects are primarily identified for the most mentioned stocks on the subreddit. Hence, given the pre-sample bias, the impact of the sentiment is only documented for a handful of stocks, which retail investors targeted.

Following our outline from the beginning of the section, secondly, the major part of the work's contribution is related to the source from which our sentiment measure originates. The role of social media is stressed throughout the thesis. The behavioural biases of retail investors are augmented by social media which not only increases access to information and opens up the discussion to the investor masses but also amplifies and reinforces the already present behavioural tendencies. The continuous repetition and overwhelmingly positive sentiment, generating a large amount of noise, manifest the fact, that retail investors ended up forming a cohort with reinforcing sentiment. As opposed to the discussion creating more disagreements, as suggested by Antweiler and Frank (2004) and Jiao et al. (2020) [8, 10], it instead fostered agreement and confirmation bias of retail investor masses which frequented the subreddit. Using social media as the magnifying glass for our examination of the sentiment and interactions between the individual investors, we suggest that it works to exacerbate noise trading and the customary proclivities of individual investors. Effectively, leading them to act in the same manner on the markets. Therefore, social media, extending beyond the behavioural sphere, could also enhance the significant impact retail investors have on the markets. Nevertheless, as noted earlier, this impact would be both time- and broad impact-wise limited. While the short-run dynamics indicate that both news and Reddit sentiment positively impact stock returns, we reveal a possible difference in the long-run relationship as both retail and news sentiment are negatively related to the returns, lending evidence for the theory

that the initially positively driven returns tend to reverse in the long run.

Finally, our evidence suggests that the effects of the retail trading frenzy were not exclusively contained in GameStop. Sampling some of the stocks most mentioned alongside it, we reveal similar dynamics between sentiment and the stocks' returns and volatility. Over short periods, GME and the other most mentioned stocks exhibit large and significant comovement across wide frequency bands, with AMC and BB being the standout co-movers. However, we fail to find evidence that the spillovers were primarily led by GME. Rather interestingly, this is further supported by VECM showing a relatively weak role of GME in the system, while other stocks appeared to drive the short-run dynamics, including AMC and BB. Universally, the role of sentiment is shown strong for all stocks across the board. Ultimately, the strongest comovement coincides with the largest price volatility, and as retail investors are drawn to these features, the sentiment intensifies, which effectively leads to spillovers to the alongside mentioned stocks. Albeit the comovement, similarly to the effect of sentiment, does not last long for most of the stocks - AMC being the notable exception. Given the revealed connection between retail investor sentiment and stock returns, we believe that subjecting this relationship to the area of asset pricing presents an interesting avenue for further exploration of the relationship, especially in relation to portfolio optimisation and the persistent deviation from the fundamentals created by the noise trader risk. However, notice that we predominantly utilised very high-frequency data, meaning that the time window in which sentiment was able to predict future returns is relatively short. Moreover, any subsequent predicted returns might be small and likely diminished or completely eradicated by transaction costs.

The noise trader theory dictates that, in principle, given the circumstances of limited arbitrage and uninformed irrational traders, individual investor sentiment can significantly affect asset prices, leading to deviation from fundamental values, which can further sustain for longer due to the former presupposition. In our work, we draw a parallel to that phenomenon using directly sourced Reddit investor sentiment from a high-frequency dataset. The extracted sentiment is demonstrated to largely reflect the behavioural biases of retail investors, backed by extensive literature, which are exacerbated by social media. In short periods, characterised by mostly large price volatility and frequent news coverage, individual investors can significantly impact the prices of the targeted stocks. However, while the impact is time limited and lacks breadth regarding the number of targeted stocks, we can observe sustained deviation from fundamentals. Moreover, the subsequent reversal of the relationship between retail investor sentiment and stock returns is evidenced by the negative long-run dynamics. News sentiment exhibits the same long-run relationship, implying that over a long investment horizon, the effect does not differ as opposed to the short-run dynamics, where news coverage was found to be largely

driven by GME.

References

[1] V. Hajric and E. Graffeo, "Retail traders slide back below 20total volume." https://www.bloomberg.com/news/articles/2021-11-17/retail-traders-retreat-as-choppy-markets-challenge-easy-profits, 2021.

[2] K. Martin and R. Wigglesworth, "Rise of the retail army: the amateur traders transforming markets." https://www.ft.com/content/7a91e3ea-b9ec-4611-9a03-a8dd3b8bddb5, March 2021.

[3] Deloitte, "The rise of newly empowered retail investors." https://www2.deloitte.com/us/en/pages/financial-services/articles/the-future-of-retail-brokerage.html, 2021.

[4] T. Odean, "Do investors trade too much?," *The American Economic Review*, vol. 89, no. 5, pp. 1279–1298, 1999.

[5] B. M. Barber and T. Odean, "Online investors: Do the slow die first?," *The Review of Financial Studies*, vol. 15, no. 2, pp. 455–487, 2002.

[6] A. Kumar and C. M. C. Lee, "Retail investor sentiment and return comovements," *The Journal of Finance*, vol. 61, no. 5, pp. 2451–2486, 2006.

[7] B. M. Barber, T. Odean, and N. Zhu, "Do retail trades move markets?," *The Review of Financial Studies*, vol. 22, no. 1, pp. 151–186, 2009.

[8] W. Antweiler and M. Z. Frank, "Is all that talk just noise? the information content of internet stock message boards," *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[9] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, "Wisdom of crowds: The value of stock opinions transmitted through social media," *The Review of Financial Studies*, vol. 27, no. 5, pp. 1367–1403, 2014.

[10] P. Jiao, A. Veiga, and A. Walther, "Social media, news media and the stock market," *Journal of Economic Behavior and Organization*, vol. 176, pp. 63–90, 2020.

[11] M. Farrell, T. C. Green, R. Jame, and S. Markov, "The democratization of investment research and the informativeness of retail investor trading," *Journal of Financial Economics*, vol. 145, no. 2, Part B, pp. 616–641, 2022.

[12] F. Black, "Noise," *The Journal of Finance*, vol. 41, no. 3, pp. 529–543, 1986.

[13] A. Shleifer and L. H. Summers, "The noise trader approach to finance," *The Journal of Economic Perspectives*, vol. 4, no. 2, pp. 19–33, 1990.

[14] B. M. Barber and T. Odean, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *The Review of Financial Studies*, vol. 21, no. 2, pp. 785–818, 2008.

[15] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," *The Journal of Finance and Data Science*, vol. 4, no. 2, pp. 120–137, 2018.

[16] R. J. Shiller, "Narrative economics," *The American Economic Review*, vol. 107, no. 4, pp. 967–1004, 2017.

[17] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[18] G. Yen and C.-f. Lee, "Efficient market hypothesis (emh): Past, present and future," *Review of Pacific Basin Financial Markets and Policies*, vol. 11, no. 02, pp. 305–329, 2008.

[19] R. J. Shiller, "From efficient markets theory to behavioral finance," *The Journal of Economic Perspectives*, vol. 17, no. 1, pp. 83–104, 2003.

[20] W. F. M. D. Bondt and R. Thaler, "Does the stock market overreact?," *The Journal of Finance*, vol. 40, no. 3, pp. 793–805, 1985.

[21] M. S. Rozeff and W. R. Kinney, "Capital market seasonality: The case of stock returns," *Journal of Financial Economics*, vol. 3, no. 4, pp. 379–402, 1976.

[22] R. Roll, "Vas ist das? the turn-of-the-year effect and the return premia of small firms," *The Journal of Portfolio Management*, vol. 9, no. 2, pp. 18–28, 1983.

[23] M. R. Reinganum, "The anomalous stock market behavior of small firms in january: Empirical tests for tax-loss selling effects," *Journal of Financial Economics*, vol. 12, no. 1, pp. 89–104, 1983.

[24] J. R. Ritter, "The buying and selling behavior of individual investors at the turn of the year," *The Journal of Finance*, vol. 43, no. 3, pp. 701–717, 1988.

[25] A. S. Kyle, "Continuous auctions and insider trading," *Econometrica*, vol. 53, no. 6, pp. 1315–1335, 1985.

[26] J. B. D. Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "The size and incidence of the losses from noise trading," *The Journal of Finance*, vol. 44, no. 3, pp. 681–696, 1989.

[27] J. B. de Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Positive feedback investment strategies and destabilizing rational speculation," *The Journal of Finance*, vol. 45, no. 2, pp. 379–395, 1990.

[28] J. B. D. Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise trader risk in financial markets," *Journal of Political Economy*, vol. 98, no. 4, pp. 703–738, 1990.

[29] R. Mehra and E. C. Prescott, "The equity premium: A puzzle," *Journal of Monetary Economics*, vol. 15, no. 2, pp. 145–161, 1985.

[30] C. M. C. Lee, A. Shleifer, and R. H. Thaler, "Investor sentiment and the closed-end fund puzzle," *The Journal of Finance*, vol. 46, no. 1, pp. 75–109, 1991.

[31] G. W. Brown, "Volatility, sentiment, and noise traders," *Financial Analysts Journal*, vol. 55, no. 2, pp. 82–90, 1999.

[32] J. N. Bodurtha, D.-S. Kim, and C. M. C. Lee, "Closed-end country funds and u.s. market sentiment," *The Review of Financial Studies*, vol. 8, no. 3, pp. 879–918, 1995.

[33] J. Pontiff, "Excess volatility and closed-end funds," *The American Economic Review*, vol. 87, no. 1, pp. 155–169, 1997.

[34] R. J. Barkham and C. W. R. Ward, "Investor sentiment and noise traders: Discount to net asset value in listed property companies in the u.k.," *The Journal of Real Estate Research*, vol. 18, no. 2, pp. 291–312, 1999.

[35] B. Swaminathan, "Time-varying expected small firm returns and closed-end fund discounts," *The Review of Financial Studies*, vol. 9, no. 3, pp. 845–887, 1996.

[36] S. Kothari and J. Shanken, "Book-to-market, dividend yield, and expected market returns: A time-series analysis," *Journal of Financial Economics*, vol. 44, no. 2, pp. 169–203, 1997.

[37] R. Neal and S. M. Wheatley, "Do measures of investor sentiment predict returns?," *The Journal of Financial and Quantitative Analysis*, vol. 33, no. 4, pp. 523–547, 1998.

[38] M. Baker and J. Wurgler, "The equity share in new issues and aggregate stock returns," *The Journal of Finance*, vol. 55, no. 5, pp. 2219–2257, 2000.

[39] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, vol. 61, no. 4, pp. 1645–1680, 2006.

[40] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *The Journal of Economic Perspectives*, vol. 21, no. 2, pp. 129–151, 2007.

[41] C. M. Lee, "Earnings news and small traders: An intraday analysis," *Journal of Accounting and Economics*, vol. 15, no. 2, pp. 265–302, 1992.

[42] B. M. Barber and T. Odean, "Trading is hazardous to your wealth: The common stock investment performance of individual investors," *The Journal of Finance*, vol. 55, no. 2, pp. 773–806, 2000.

[43] B. M. Barber and T. Odean, "Boys will be boys: Gender, overconfidence, and common stock investment," *The Quarterly Journal of Economics*, vol. 116, no. 1, pp. 261–292, 2001.

[44] C.-C. Chang, P.-F. Hsieh, and Y.-H. Wang, "Sophistication, sentiment, and misreaction," *The Journal of Financial and Quantitative Analysis*, vol. 50, no. 4, pp. 903–928, 2015.

[45] A. Jackson, "The aggregate behaviour of individual investors," *SSRN Electronic Journal*, 2003.

[46] L. Feng and M. S. Seasholes, "Correlated trading and location," *The Journal of Finance*, vol. 59, no. 5, pp. 2117–2144, 2004.

[47] N. Barberis, A. Shleifer, and J. Wurgler, "Comovement," *Journal of Financial Economics*, vol. 75, no. 2, pp. 283–317, 2005.

[48] R. Kaniel, G. Saar, and S. Titman, "Individual investor trading and stock returns," *The Journal of Finance*, vol. 63, no. 1, pp. 273–310, 2008.

[49] T. FOUCAULT, D. SRAER, and D. J. THESMAR, "Individual investors and volatility," *The Journal of Finance*, vol. 66, no. 4, pp. 1369–1406, 2011.

[50] S. C. Andrade, C. Chang, and M. S. Seasholes, "Trading imbalances, predictable reversals, and cross-stock price pressure," *Journal of Financial Economics*, vol. 88, no. 2, pp. 406–423, 2008.

[51] M. W. Brandt, A. Brav, J. R. Graham, and A. Kumar, "The idiosyncratic volatility puzzle: Time trend or speculative episodes?," *The Review of Financial Studies*, vol. 23, no. 2, pp. 863–899, 2010.

[52] M. J. Kamstra, L. A. Kramer, and M. D. Levi, "Winter blues: A sad stock market cycle," *The American Economic Review*, vol. 93, no. 1, pp. 324–343, 2003.

[53] A. Edmans, D. García, and Ø. Norli, "Sports sentiment and stock returns," *The Journal of Finance*, vol. 62, no. 4, pp. 1967–1998, 2007.

[54] W. N. Goetzmann, D. Kim, A. Kumar, and Q. Wang, "Weather-induced mood, institutional investors, and stock returns," *The Review of Financial Studies*, vol. 28, no. 1, pp. 73–111, 2015.

[55] A. Edmans, A. Fernandez-Perez, A. Garel, and I. Indriawan, "Music sentiment and stock returns around the world," *Journal of Financial Economics*, 2021.

[56] D. M. Cutler, J. M. Poterba, and L. H. Summers, "What moves stock prices?," *The Journal of Portfolio Management*, vol. 15, no. 3, pp. 4–12, 1989.

[57] G. Gidofalvi and C. Elkan, "Using news articles to predict stock price movements," *Department of Computer Science and Engineering, University of California, San Diego*, 2001.

[58] G. Fung, J. Yu, and H. Lu, "The predicting power of textual information on financial markets.," *IEEE Intelligent Informatics Bulletin*, vol. 5, pp. 1–10, 01 2005.

[59] P. C. Tetlock, "All the news that's fit to reprint: Do investors react to stale information?," *The Review of Financial Studies*, vol. 24, no. 5, pp. 1481–1512, 2011.

[60] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," *ACM Trans. Inf. Syst.*, vol. 27, mar 2009.

[61] Z. DA, J. ENGELBERG, and P. GAO, "In search of attention," *The Journal of Finance*, vol. 66, no. 5, pp. 1461–1499, 2011.

[62] Z. Da, J. Engelberg, and P. Gao, "The Sum of All FEARS Investor Sentiment and Asset Prices," *The Review of Financial Studies*, vol. 28, pp. 1–32, 10 2014.

[63] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Scientific Reports*, vol. 3, no. 1, p. 1684, 2013.

[64] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, "Quantifying wikipedia usage patterns before stock market moves," *Scientific Reports*, vol. 3, no. 1, p. 1801, 2013.

[65] D. Bradley, J. H. Jr., R. Jame, and Z. Xiao, "Place your bets? the market consequences of investment advice on reddit's wallstreetbets," *SSRN Electronic Journal*, 2021.

[66] Z. Umar, O. B. Adekoya, J. A. Oliyide, and M. Gubareva, "Media sentiment and short stocks performance during a systemic crisis," *International Review of Financial Analysis*, vol. 78, p. 101896, 2021.

[67] Z. Umar, M. Gubareva, I. Yousaf, and S. Ali, "A tale of company fundamentals vs sentiment driven pricing: The case of gamestop," *Journal of Behavioral and Experimental Finance*, vol. 30, p. 100501, 2021.

[68] Z. Umar, I. Yousaf, and A. Zaremba, "Comovements between heavily shorted stocks during a market squeeze: Lessons from the gamestop trading frenzy," *Research in International Business and Finance*, vol. 58, p. 101453, 2021.

[69] C. Long, B. M. Lucey, and L. Yarovaya, ""i just like the stock" versus "fear and loathing on main street" : The role of reddit sentiment in the GameStop short squeeze," *SSRN Electronic Journal*, 2021.

[70] N. Fusari, R. Jarrow, and S. Lamichhane, "Testing for asset price bubbles using options data," *SSRN Electronic Journal*, p. 51, 11 2020.

[71] C. M. Jones, A. V. Reed, and W. Waller, "When brokerages restrict retail investors, does the game stop?," *SSRN Electronic Journal*, p. 79, 11 2021.

[72] M. P. Franklin Allen, Eric Nowak and A. Tengulov, "Squeezing shorts through social news platforms," *SSRN Electronic Journal*, p. 73, 8 2021.

[73] G. Baltussen, Z. Da, S. Lammers, and M. Martens, "Hedging demand and market intraday momentum," *Journal of Financial Economics*, vol. 142, no. 1, pp. 377–403, 2021.

[74] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," 2014.

[75] P. Kumar and E. Foufoula-Georgiou, "Wavelet analysis for geophysical application," *Reviews of Geophysics*, vol. 35, pp. 385–412, 11 1997.

[76] R. Kronland-Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music Journal*, vol. 12, no. 4, pp. 11–20, 1988.

[77] J. B. Ramsey and C. Lampart, "Decomposition of economic relationships by timescale using wavelets," *Macroeconomic Dynamics*, vol. 2, no. 1, pp. 49–71, 1998.

[78] P. M. Crowley, "A guide to wavelets for economists," *Journal of Economic Surveys*, vol. 21, no. 2, pp. 207–267, 2007.

[79] A. Rua and L. C. Nunes, "International comovement of stock market returns: A wavelet analysis," *Journal of Empirical Finance*, vol. 16, no. 4, pp. 632–639, 2009.

[80] M. Gallegati, "A wavelet-based approach to test for financial market contagion," *Computational Statistics and Data Analysis - CS DA*, vol. 56, 11 2012.

[81] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61 – 78, 1998.

[82] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes in Geophysics*, vol. 11, pp. 561–566, Nov. 2004.

[83] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987.

[84] D. F. Hendry and K. Juselius, "Explaining cointegration analysis: Part ii," *The Energy Journal*, vol. 22, no. 1, pp. 75–120, 2001.

[85] O. Filip, K. Janda, L. Kristoufek, and D. Zilberman, "Food versus fuel: An updated and expanded evidence," *Energy Economics*, vol. 82, pp. 152–166, 2019. Replication in Energy Economics.

[86] S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models," *Econometrica*, vol. 59, no. 6, pp. 1551–1580, 1991.

[87] S. Johansen, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, 1995.

[88] E. Martina, E. Rodriguez, R. Escarela-Perez, and J. Alvarez-Ramirez, "Multiscale entropy analysis of crude oil price dynamics," *Energy Economics*, vol. 33, no. 5, pp. 936–947, 2011.

[89] L. Vacha and J. Barunik, "Co-movement of energy commodities revisited: Evidence from wavelet coherence analysis," *Energy Economics*, vol. 34, no. 1, pp. 241–247, 2012.

[90] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979.

[91] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?," *Journal of Econometrics*, vol. 54, no. 1, pp. 159–178, 1992.

[92] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 2022/07/09/ 1978.

[93] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, December 1974.

[94] L. Ljung, *System Identification*, pp. 163–173. Boston, MA: Birkhäuser Boston, 1998.

[95] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[96] L. H. Pedersen, "Game on: Social networks and markets," *SSRN Electronic Journal*, p. 52, 10 2021.

[97] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

Appendix A

Fig. A.1: Stock returns and sentiment

In addition to the respective log-returns sentiment, both aggregation sum and mean of Reddit sentiment scores are supplied. The original sample covers almost an entire year from January 2021 but only until November 2021. Note that the y-axis do not correspond to each other, for better visibility.

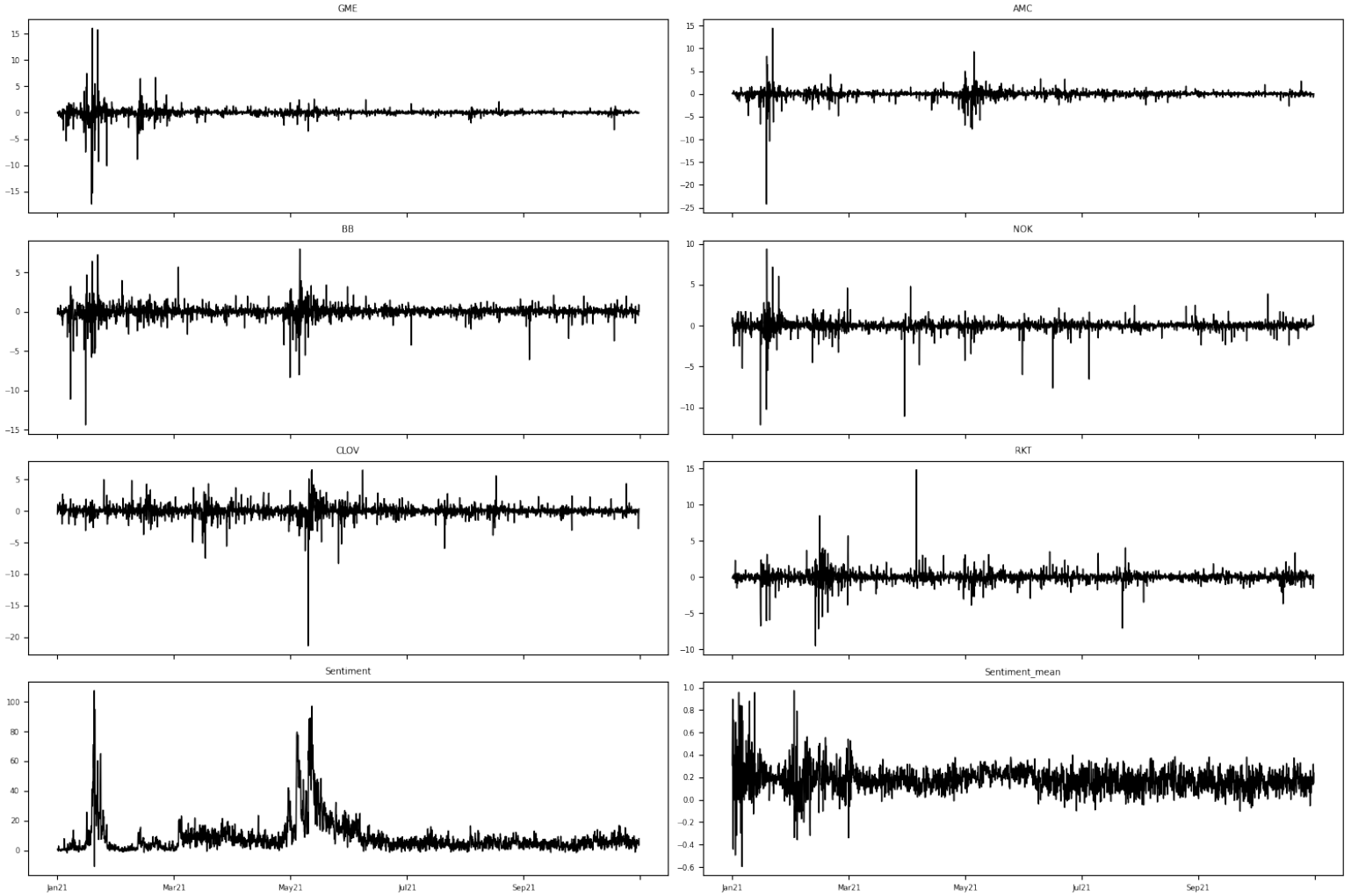


Fig. A.2: Difference in sampling frequency

Density plots for the mean of Reddit sentiment aggregated at different frequencies, namely 30 minutes, 2 and 4 hours, 1 and 5 trading days.

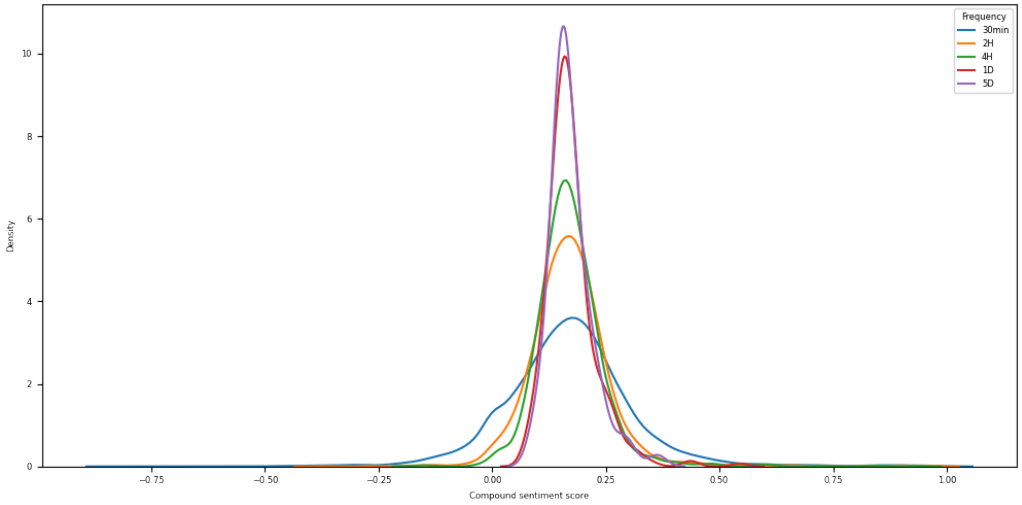
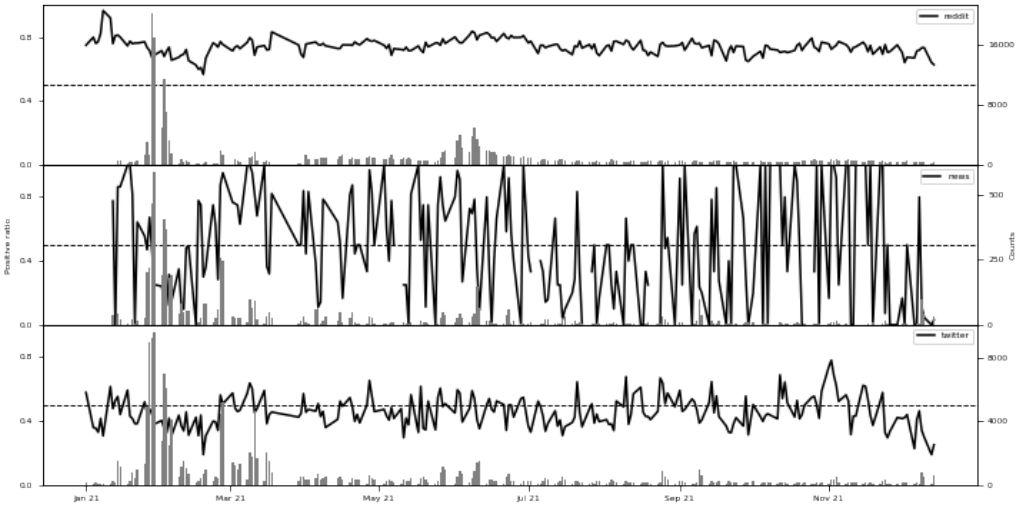


Fig. A.3: Sentiment comparison

Comparison between Reddit, news and Twitter sentiment using the share of positively scored submissions/stories/tweets. Neutral scores are omitted. The bars represent the overall counts (RHS), excluding neutral posts, which measure activity. The dashed line is drawn at 0.5, i.e. line above indicates more than 50% of all scored posts were positive.



Tab. A.1: Granger causality
p-values of Granger causality tests for short-run dynamics for Reddit sentiment and the sampled stocks. The tested null hypothesis is represented as *the variable in the row does not Granger cause the variable in the column*. Daily aggregated data on Reddit sentiment, GME, AMC, BB, NOK, CLOV and RKT log-returns.

	GME	*Sent*	AMC	BB	NOK	CLOV	RKT
GME		<0.01	<0.01	0.0100	<0.01	0.1915	0.1486
Sentiment	<0.01		0.0144	0.4859	0.0456	0.5873	0.1811
AMC	0.2485	<0.01		<0.01	<0.01	0.7917	0.3595
BB	0.8029	<0.01	0.0751		0.0202	0.1350	0.0891
NOK	<0.01	<0.01	<0.01	<0.01		0.5324	0.1534
CLOV	0.8041	0.7018	0.9401	0.9691	0.1551		0.0037
RKT	0.5844	0.8456	0.8523	0.7866	0.6733	0.8169	

Appendix B

The code repository containing the used data and scripts used to scrape the Reddit data as well as some of the financial is available on the following address: `https://github.com/kevintrng/MasThe/sentiment`. Most of the coding exercises were done in Python, including the graphical outputs. The Bloomberg data and white paper on the news analytics function of the Bloomberg Terminal can be provided upon request.

Fig. A.4: Volatility spillovers

Wavelet coherence of AMC, BB, NOK, CLOV and RKT with GME volatility. Based on intraday 30-min data and 10-interval rolling standard deviation. Vertical axis runs from 1 to 128 hours, i.e. up to 16 trading days. Thin white contours denote 5% significance level against red noise. The arrows represent relative phase difference. Cone of influence (COI) is delineated by the white blurred area where edge effects take place.