Economics Letters 244 (2024) 112027

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Vocabulary Herfindahl Index (VocaHIn): Linguistic dominance and collective effervescence in WallStreetBets

Ko Hayakawa [a], Yoichi Otsubo [b,c,*], Ser-huang Poon [c], Siliang Wei [c]

[a] The University of Tokyo, Research Center for Advanced Science and Technology, Tokyo, Japan
[b] Kobe University, Graduate School of Economics, Kobe, Japan
[c] The University of Manchester, Alliance Manchester Business School, Manchester, United Kingdom

ARTICLE INFO

JEL classification code:
G1

Keywords:
Vocabulary Herfindahl Index (VocaHIn)
Collective effervescence
Retail investors
Meme stocks

ABSTRACT

Our analysis of over 150,000 WallStreetBets comments reveals dominant phrases when members of the community experience collective excitement and solidarity (*Collective Effervescence*). Using the novel *Vocabulary Herfindahl Index* (*VocaHIn*), we find that stock returns volatility increases after *CE*, and vice versa.

1. Introduction

During the beginning of 2021, GameStop's stock (GME) experienced a tenfold surge within a week. This abrupt change in the price, characteristic of *meme* stocks, is often attributed to the collective action of retail investors on the Reddit discussion board WallStreetBets (WSB). As "zero-commission" trading apps like Robinhood gained popularity, which might fuel the *meme* stock episode, academic research in finance explored the link between online retail investors' social media communication and financial markets. Pedersen (2022) developed a theoretical model explaining *meme* stock bubbles, emphasizing the contagion of beliefs among diverse investor types. Gendron et al. (2023) examined WSB users' unique communication styles in detail. Long et al. (2023) investigates the influence of WSB discussions on GME price dynamics, utilizing a sentiment analysis to explore correlations with short-term stock returns. Bradley et al. (2024) analysed the informativeness of the WSB posts and concludes that the GME event altered WSB's culture, adversely affecting retail investors.

In this paper, we shed light on a distinctive feature of the phenomenon surrounding *meme* stock bubbles by focusing on the social aspect of its driving force, *Collective Effervescence* (*CE*), a state where individuals experience collective excitements and *solidarity* (Durkheim, 1912). Trading *meme* stocks and sharing experiences through commenting on WSB fosters *CE*, thereby enhancing the cohesion of the online community (Hayakawa and Otsubo, 2023). *CE* in online community involves participants creating a fictitious sense of physical togetherness through unique expressions and shared jargon (Hayakawa and Ide, 2009). The paper delves into cases of *CE* within the WSB, focusing on its prominent user, DeepF***ingValue (DFV).

Our proposed metric, the *VocaHIn,* serves as an innovative tool for detection of *CE*. Grounded in the observation that a concentrated set of phrases dominates comments during *CE* events, *VocaHIn* quantifies such dominance. The index peaked during GME's most turbulent period, signalling strengthened *solidarity* within the WSB community amidst the uncertainty. The relation between *VocaHIn* and realized volatility revealed from our regression analyses underscores the link between linguistic patterns and market dynamics.

2. Collective effervescence

A striking example of *CE* can be observed in DFV's post on January 28th, 2021. On that day, GME share price plummeted from the previous day's close of \$347.51 to \$193.60 (i.e. −58.50%). This represented the second largest daily loss the stock had ever experienced (See Fig. 1, Panels A and B). DFV himself incurred a loss of \$14.8 M in just one day. Despite this immense setback, he chose not to sell his position.[1] This update became the most popular post, garnering over 23 thousand comments.

* Corresponding author at: Kobe University, Graduate School of Economics, 2-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo Prefecture, 657-8501, Japan.
E-mail address: otsubo@econ.kobe-u.ac.jp (Y. Otsubo).

[1] This can be confirmed from the screenshot of his account on his January 27th, 2021 and January 28th, 2021 posts. See Fig. 2 Panels A and B.

https://doi.org/10.1016/j.econlet.2024.112027
Received 14 March 2024; Received in revised form 10 October 2024; Accepted 12 October 2024
Available online 13 October 2024

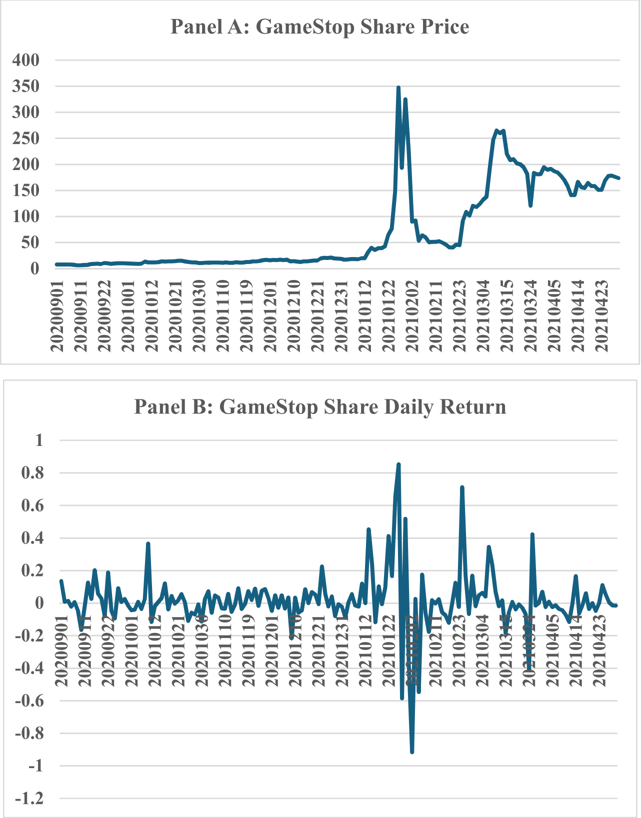


Fig. 1. GameStop daily share price and returns.
Notes: Panel A plots the end of the day closing prices of GameStop and Panel B plots the daily returns from 1st September 2020 to 29th April 2021.

The community's reactions to the post vividly showcased *CE*. In the example (Fig. 3), members heavily repeated the *meme* phrase "If he's still in, I'm still in". WSB members experience *solidarity* through repeating these *meme* phrases. Such phenomenon has been reported in Hayakawa and Ide (2009). The participants of Japanese discussion board create a sense of togetherness through peculiar way of expression while sharing their internally circulated jargon and further identify a shareable narrative.

Fig. 4 Panel A counts comments containing solely the phrase "If he is still in, I am still in."[2] Our analysis of these comments confirms that the phrase "If he is still in, I am still in" was used over one thousand times on 28th January 2021, coinciding with the observed *CE*. Furthermore, we observed that the phrase had not been widespread prior to this date and was suddenly used in an excessive manner on 28th January 2021.

Regarding the timing of *CE* in connection with market conditions, we find that it coincided with the most volatile period of GME's share price. There are some other phrases commonly used by WSB users during their collective excitement, such as "This is the way", "Legend", and "We like the stock."[3] Panel B of Fig. 4 plots the number of comments solely composed of one of these three common phrases. The phrase "This is the way" had been the most popular phrase, being used more frequently than "If he is still in, I am still in" until 27th January 2021, a day before DFV experienced significant loss and WSB users experienced *CE*.

3. Vocabulary Herfindahl Index (VocaHIn)

We propose a metric for detecting *CE* without relying on manual inspection of comments. We characterise *CE* from the observation that a small group of phrases dominating the comments, leading to a reduction in the variety of phrases and vocabulary used. Leveraging this characteristic, we introduce an index that measures the dominance of popular phrase usage similar to the Herfindahl index, commonly used to measure market share dominance. We refer to this measure as the Vocabulary Herfindahl Index (*VocaHIn*).[4] For each of DFV's posts, we count the number of times an exactly identical comment, referred to as a "phrase," is posted. We then rank the phrases by their frequency of appearance for each post.[5] *VocaHIn* is calculated as follows:

$$VocaHIn_t = \sum_{i}^{N} p_{i,t}^2$$

where $p_{i,t}$ is the proportion of the top *i*th phrase used among the top *N* phrases in DFV's post on day *t*. For each of the DFV's posts, we calculate $VocaHIn_t$ with the twenty most used phrases (i.e., $N = 20$). A higher value of the index indicates the dominance of limited number of phrases, thus implying the occurrence of *CE*.

In Fig. 5 Panel A, we present the time series of $VocaHIn_t$. The metric peaks during the most volatile period as measured by realized volatility. In Panel B we also plot $VocaHIn_t$ alongside absolute daily returns, confirming the relationship between *CE* and the volatility of the GME share prices. The correlation coefficient between $VocaHIn_t$ and the realized volatility is 0.694 (0.4793 with the absolute daily returns). In the following section we further examine the relationship.

4. Relationship between *CE* and GameStop share price volatility

To establish a relationship between WSB users' *CE* and GameStop share price volatility, we exploit the timing of DFV's posts, which are always outside regular trading hours. We estimate two regression models. The first tests the impact of *CE* on volatility, and the second examines the effect of volatility on *CE*:

$$Volatility_t = \beta_0 + \beta_1 VocaHIn_{t^-} + \beta_2 Volume_t + \beta_3 Short\ Sale_t + \beta_4 Retail_t + \varepsilon_{1,t} \quad (1)$$

$$VocaHIn_{t^+} = \beta_0 + \beta_1 Volatility_t + \beta_2 Volume_t + \beta_3 Short\ Sale_t + \beta_4 Retail_t + \varepsilon_{2,t} \quad (2)$$

Eq. (1) estimates how *CE* among WSB users affects the volatility of GameStop's share price. Here, $VocaHIn_{t^-}$ measures the *CE* among WSB users, while $Volatility_t$ represents either the absolute daily return (*Abs return*) or the daily realized volatility calculated from 5-minute returns (*RV5min*), and $\varepsilon_{1,t}$ is the error term. As DFV always submits his posts outside regular trading hours, to estimate the impact of WSB users' *CE* on GameStop share price volatility, the volatility is calculated from the closest trading hours *following* the time of DFV's post. In other words, $VocaHIn_{t^-}$ precedes the regular trading hours of day *t*, hence the minus superscript, t^-, indicating this timing. As we focus on the occurrence of

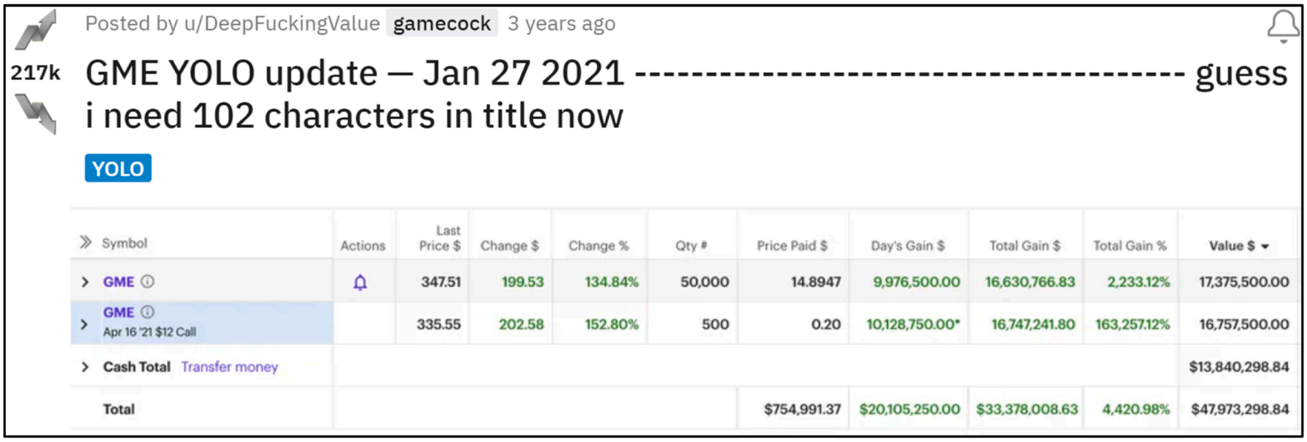
[2] We focus on the specific expression "If he is still in, I am still in," without considering variations such as "If he's in, I'm in." Including these variants in our counts would not substantially alter the overall usage trend of the phrase.

[3] These phrases are respectively the 2nd, 3rd, and 5th most posted phrases (The 4th most posted is "If hes still in, I am still in," with the typo of missing apostrophe between 'he' and 's'.). Note that "If he's still in, I'm still in" ('he's' and 'I'm' with an apostrophe) and "If he is still in, I am still in" ('he is' and 'I am' without using an apostrophe) are treated as identical comments.

[4] "*voca-hin*" (ボキャ貧) is a Japanese slang term describing a person who only uses a limited variety of vocabulary in their speech, discussion, or comments.

[5] For example, in Fig. 4, Panel A, we report the counts of the exact phrase, "If he's still in, I'm still in." Comments that are variants of this phrase, such as "If he's in, I'm in," or versions with typos like "If hes still in, Im still in," or with additional phrases like "If he's still in, I'm still in. GME to the moon!" are not included in our counts. We acknowledge that this approach may underestimate the usage of comments that convey the same meaning as "If he's still in, I'm still in."

Panel A



Posted by u/DeepFuckingValue gamecock 3 years ago

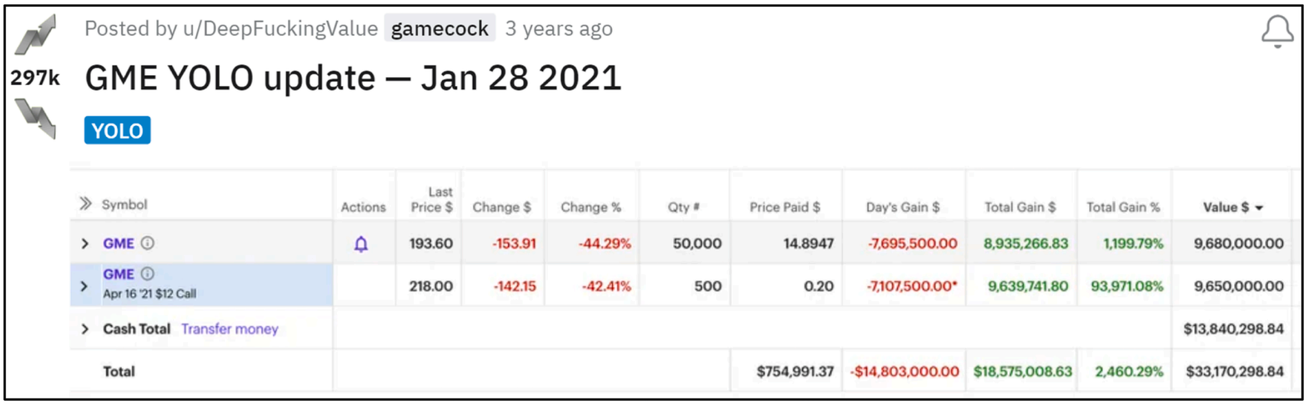
217k

GME YOLO update — Jan 27 2021 ------------------------------------ guess i need 102 characters in title now

YOLO

Symbol	Actions	Last Price $	Change $	Change %	Qty #	Price Paid $	Day's Gain $	Total Gain $	Total Gain %	Value $
GME	🔔	347.51	199.53	134.84%	50,000	14.8947	9,976,500.00	16,630,766.83	2,233.12%	17,375,500.00
GME Apr 16 '21 $12 Call		335.55	202.58	152.80%	500	0.20	10,128,750.00*	16,747,241.80	163,257.12%	16,757,500.00
Cash Total Transfer money										$13,840,298.84
Total						$754,991.37	$20,105,250.00	$33,378,008.63	4,420.98%	$47,973,298.84

Panel B



Posted by u/DeepFuckingValue gamecock 3 years ago

297k

GME YOLO update — Jan 28 2021

YOLO

Symbol	Actions	Last Price $	Change $	Change %	Qty #	Price Paid $	Day's Gain $	Total Gain $	Total Gain %	Value $
GME	🔔	193.60	-153.91	-44.29%	50,000	14.8947	-7,695,500.00	8,935,266.83	1,199.79%	9,680,000.00
GME Apr 16 '21 $12 Call		218.00	-142.15	-42.41%	500	0.20	-7,107,500.00*	9,639,741.80	93,971.08%	9,650,000.00
Cash Total Transfer money										$13,840,298.84
Total						$754,991.37	-$14,803,000.00	$18,575,008.63	2,460.29%	$33,170,298.84

Fig. 2. DFV's post on 27th and 28th January 2021.
Panel A: DFV post on 27. Jan. 2021. url: https://www.reddit.com/r/wallstreetbets/comments/l6ekdz/gme_yolo_update_jan_27_2021_guess_i_need_102/.
Panel B: DFV post on 28. Jan. 2021. url: https://www.reddit.com/r/wallstreetbets/comments/l78uct/gme_yolo_update_jan_28_2021/.
Notes: The screenshots capture DFV's posts on the WallStreetBets on January 27th and 28th, 2021. In these posts, DFV discloses his account information and his positions, including gains and losses, with the community. The figures listed under the "Qty #" column in both posts suggest that DFV maintained his position despite experiencing a 44.29% loss in GameStop shares and a 42.41% loss in call options.

CE in DFV's posts, the number of observations corresponds to the number of posts made by him, totalling 58.

RV5min is calculated using a standard approach by summing the squared price changes during trading hours, from 9am to 4:30pm. To address potential microstructure noise resulting from bid-ask bounce, we utilize the best bid and ask quotes to calculate mid-quotes for return calculations. The formula is $\sqrt{252\sum_{n=1}^{N} r_{t,n}^2}$, where N denotes the total number of intraday returns, and $r_{t,n}$ denotes the corresponding 5-minute mid-quote return on day t. The data is sourced from LSEG DataScope.

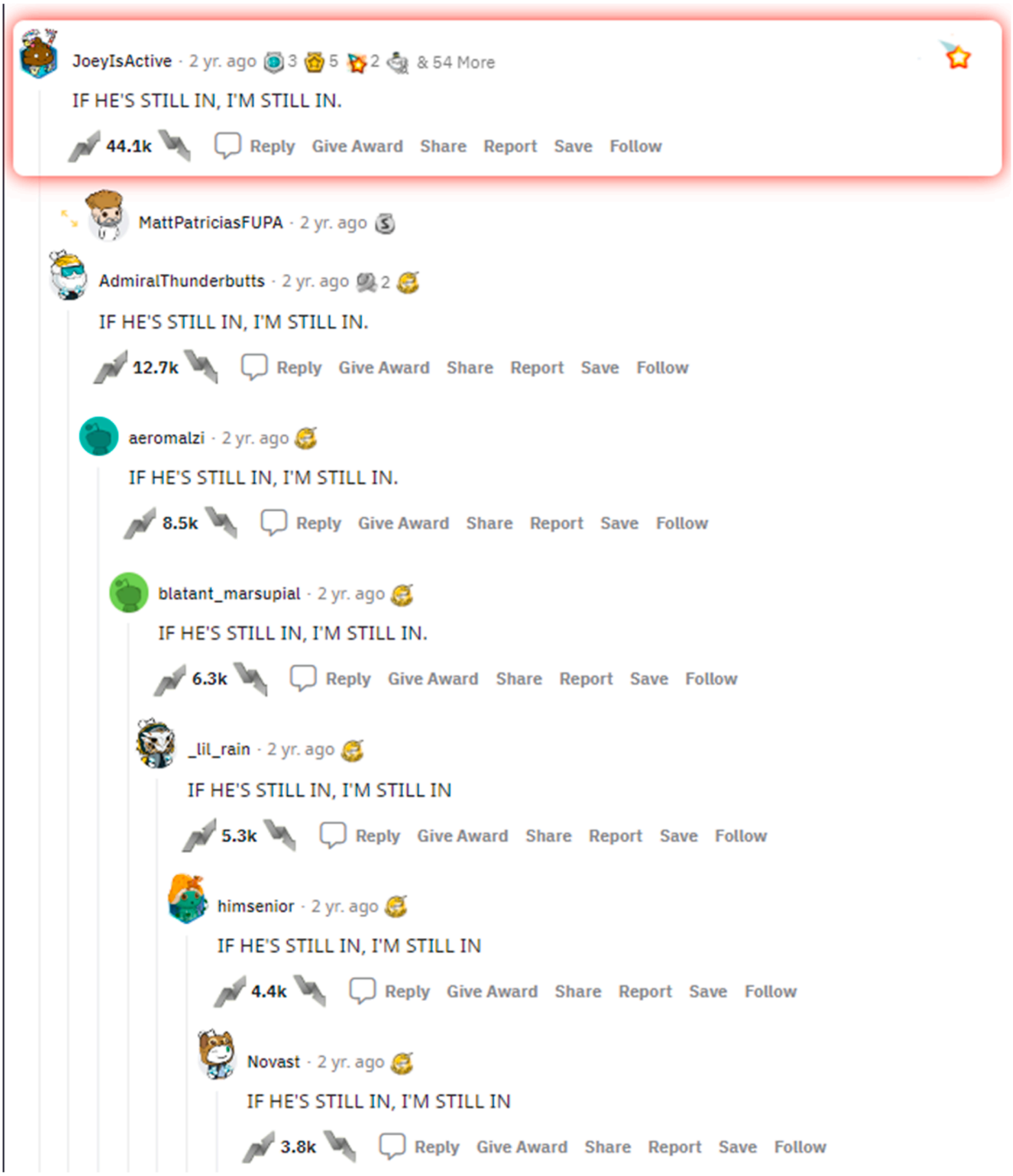
As control variables, we include trading volume variables that affect volatility. In addition to the standard daily trading volume, $Volume_t$, we incorporate two particularly relevant variables. The first variable, $Short\ Sale_t$, measures the proportion of short sale volume over total trading volume. The variable's impact is expected to be significant as short sellers were targeted by Redditors, and the potential occurrence of a short squeeze were anticipated (Long et al., 2023). Higher short selling activities could increase uncertainty and thus lead to higher volatility in the stock price. The short sale data is obtained from Financial Industry Regulatory Authority (FINRA) and Chicago Board of Exchange (CBOE).[6]

The second variable, $Retail_t$, measures the proportion of retail trading volume over total trading volume. Retail trading volume is proxied by retail marketable orders, calculated using the method introduced by Boehmer, Jones, Zhang, and Zhang (2021).[7]

Eq. (2) estimates the relationship between the volatility of GameStop's share price and the *CE* among WSB users employing $VocaHIn_{t^+}$ as the explained variable of the regression. In contrast to Eq. (1), the volatility and volume variables in Eq. (2) are measured from the closest trading hours *preceding* the time of DFV's post to estimate the effect of volatility on *CE*. In other words, $VocaHIn_{t^+}$ succeeds the regular trading

[6] FINRA is a widely used source of short sale data in finance literature (e.g., Wu et al. (2024), among others). It mainly reports NYSE and NASDAQ short sales. To gain a more comprehensive understanding of short selling activities, we supplement this data with short selling volume information from the CBOE.

[7] Following Boehmer et al. (2021), we identify retail trading volume using sub-penny price improvements in trade and quote data from LSEG DataScope. Retail sell transactions are identified by execution prices with a sub-penny portion between 0.006 and 0.010, while retail buy transactions are identified by execution prices with a sub-penny portion between 0.001 and 0.040. We then aggregate the total marketable retail buy and sell volumes.



JoeyIsActive · 2 yr. ago 3 5 2 & 54 More

IF HE'S STILL IN, I'M STILL IN.

44.1k Reply Give Award Share Report Save Follow

MattPatriciasFUPA · 2 yr. ago

AdmiralThunderbutts · 2 yr. ago 2

IF HE'S STILL IN, I'M STILL IN.

12.7k Reply Give Award Share Report Save Follow

aeromalzi · 2 yr. ago

IF HE'S STILL IN, I'M STILL IN.

8.5k Reply Give Award Share Report Save Follow

blatant_marsupial · 2 yr. ago

IF HE'S STILL IN, I'M STILL IN.

6.3k Reply Give Award Share Report Save Follow

_lil_rain · 2 yr. ago

IF HE'S STILL IN, I'M STILL IN

5.3k Reply Give Award Share Report Save Follow

himsenior · 2 yr. ago

IF HE'S STILL IN, I'M STILL IN

4.4k Reply Give Award Share Report Save Follow

Novast · 2 yr. ago

IF HE'S STILL IN, I'M STILL IN

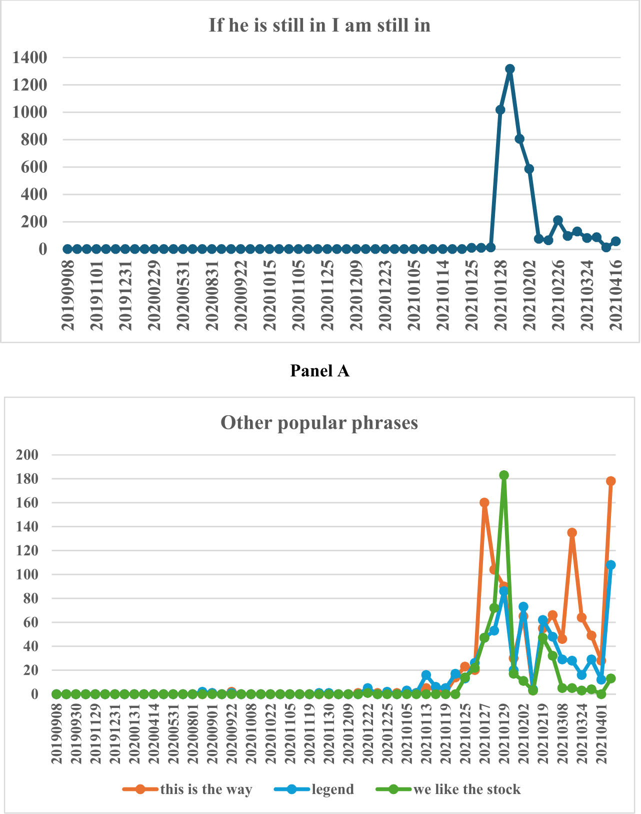
3.8k Reply Give Award Share Report Save Follow

Fig. 3. Example of Collective Effervescence.
Notes: Comments reacting to the DFV's post on 28th January 2021.

hours of day t, hence the plus superscript, t^+, indicating this timing. In addition to the *Abs. return* and *RV5min*, we also employ *RV 20days* which is realized volatility calculated from the daily returns of 20 trading days preceding DFV's post. $\varepsilon_{2,t}$ is the error term. The volume variables are included as control variables because they can influence the collective excitement of WSB users. Increased short-selling activities could affect *VocaHIn*, potentially through anger, while a surge in retail trading activities might encourage and trigger posts by similarly excited WSB users.[8]

Table 1 presents the results of Eq. (1). We also re-run the regression by adding a lagged volatility measure to Eq. (1) as volatility tends to be persistent. The results indicate that *CE*, as measured by *VocaHIn*, has a significant positive impact on both absolute daily returns and *RV5min*. This suggests that higher levels of *CE* among WSB users are associated

[8] A notable episode highlighting the tension between short-sellers and WSB users involved a tweet by the short-seller Citron Research and the subsequent reaction from WSB users. This incident received a wide media coverage, for instance: https://markets.businessinsider.com/news/stocks/citron-research-livestream-gamestop-position-halted-twitter-hack-attempts-2021-1-1029993391



Panel A

Panel B

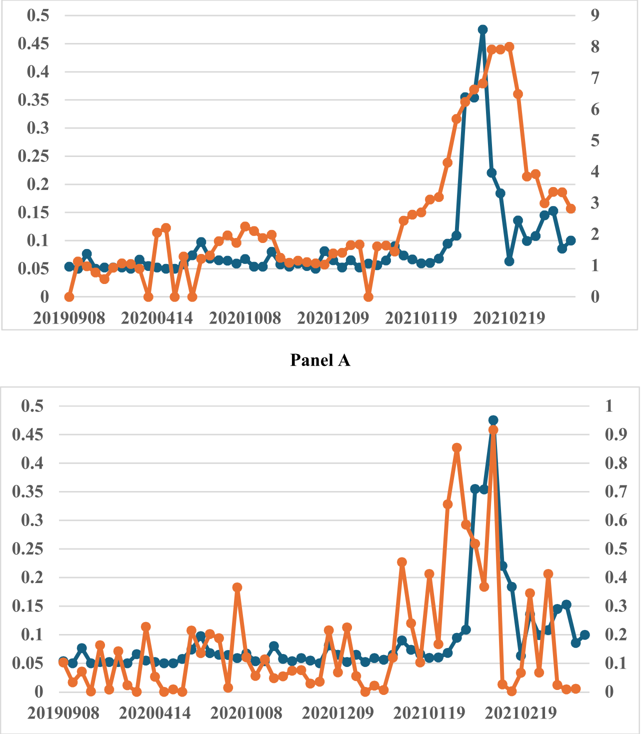
Fig. 4. Number of Comments.
Notes: Panel A plots the number of comments solely composed by the phrase “If he is still in, I am still in”. Panel B plots the number of comments to DFV’s post only consisting of the phrases “This is the way”, “Legend”, “We like the stock.”.

with increased volatility in GameStop’s share price. The results of Eq. (2) are reported in Table 2. Historical volatility (RV 20 days), short-term volatility (*RV5min*), and absolute daily returns consistently influence *CE* among WSB users. In sum, we find an upsurge in stock returns volatility following *CE* and similarly, *CE* tends to intensify after experiencing a volatile market.

We also find that the higher trading volume contributes to higher volatility. In contrast, short selling activity does not have a clear impact on volatility in this context. Retail trading activity is associated with lower volatility. Trading volume, short sale, and retail trading do not impact *CE*. These are interesting findings, but further analysis is beyond the scope of this paper.

5. Conclusion

By introducing *VocaHIn*, this paper demonstrates a significant relationship between dominant phrases in WSB discussions and GameStop’s share price volatility. Our study has limitations including a focus on DFV’s posts and a specific period. Expanding the dataset to include other influencers’ posts, meme stocks, and social media platforms could enhance our understanding of *CE*. Establishing causality between *CE* and volatility remains challenging, necessitating future research using natural experiments. Future studies could explore the impact of other platforms such as X, Stocktwits, and eToro. Developing predictive models using *VocaHIn* and other metrics may help anticipate high volatility periods driven by *CE* in social media communities. These models would have practical applications for regulators and investors in general to identify volatility spikes and enable portfolio managers to assess *meme* stock risks and adjust risk management strategies accordingly.



Panel A

Panel B

Fig 5. VocaHin and GameStop Share price Volatility.
Notes: This figure plots VocaHin with the volatility measures. Panel A plots VocaHIn and realized volatility. The red line plots the realized volatility (right axis) while the blue line plots the VocaHIn (left axis). Panel B plots VocaHIn and daily absolute return. The red line plots the absolute return (right axis) while the blue line plots the VocaHIn (left axis).

Table 1
Impact of CE on Volatility.

	Abs. return (1)	*RV5min* (2)	Abs. return (3)	*RV5min* (4)
Constant	−1.741*** (0.306)	−14.209*** (2.502)	−1.850*** (0.355)	−9.591*** (2.838)
VocaHIn	0.788*** (0.248)	6.964*** (2.031)	0.822*** (0.256)	3.908* (2.177)
Volume	0.133*** (0.020)	1.078*** (0.164)	0.141*** (0.024)	0.731*** (0.195)
Short Sale volume ratio	0.021 (0.159)	−0.700 (1.299)	0.018 (0.160)	−0.273 (1.226)
Retail Volume ratio	−0.982*** (0.309)	−8.121*** (2.529)	−1.037*** (0.323)	−5.939** (2.487)
Abs. return $_{t-1}$			−0.070 (0.114)	
RV5min $_{t-1}$				0.351*** (0.121)
Observations	58	58	58	58
Adj. R^2	0.646	0.657	0.641	0.699

Notes: This table presents the regression results of Eq. (1). The variables of interest is VocaHIn. Control variables include Volume, Short sale, Retail volume. (3) and (4) include lagged volatility, Abs. return and *RV5min* respectively. Standard errors are reported in parentheses. ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Chat-GPT 4o in order to edit and reduce word counts. After using this tool/service, the

Table 2
Impact of Volatility on CE.

	VocaHIn	VocaHIn	VocaHIn
	(1)	(2)	(3)
Constant	−0.014	0.342*	0.040
	(0.201)	(0.177)	(0.141)
Absolute return	0.159***		
	(0.066)		
RV5min		0.038***	
		(0.007)	
RV 20 days			0.028***
			(0.005)
Volume	0.010	−0.019	0.002
	(0.013)	(0.012)	(0.009)
Short Sale volume ratio	0.004	0.047	0.015
	(0.092)	(0.077)	(0.075)
Retail Volume ratio	−0.224	0.013	−0.142
	(0.162)	(0.144)	(0.130)
Observations	58	58	58
Adj. R^2	0.212	0.444	0.466

Notes: This table presents the regression results of Eq. (2). The variables of interests are RV 20days, Absolute return, and *RV5min*. Control variables include Volume, Short sale, Retail volume. Standard errors are reported in parentheses. ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgement

The authors deeply appreciate the invaluable financial assistance offered by the Alliance Manchester Business School and the Suntory Foundation.

Data availability

Data will be made available on request.

References

Boehmer, E., Jones, C.M., Zhang, X., Zhang, X., 2021. Tracking Retail Investor Activity. J. Finance 76, 2249–2305. https://doi.org/10.1111/jofi.13033.

Bradley, D., Hanousek Jr., J., Jame, R., Xiao, Z, 2024. Place Your Bets? The Market Consequences of Investment Research on Reddit's Wallstreetbets. Rev. Financ. Stud. 37 (5), 1409–1459. https://doi.org/10.1093/rfs/hhad098.

Durkheim, E., 1912. The Elementary Forms of Religious Life. Hollen street press.

Gendron, Y., Madelaine, A., Paugam, L., Stolowy, H., 2023. Shaping collective action in financial markets through the development of popular expertise: an Analysis of Due Diligence Posts on WallStreetBets. Available at SSRN. https://ssrn.com/abstract=4234609.

Hayakawa, K., Ide, R., 2009. Language practice on 2-Channel of Internet Board and a Sense of Community (Japanese). In: Miyake, K., et al. (Eds.), Media and Language 4. Hitsuji Syobou, Tokyo, pp. 192–219.

Hayakawa, K., Otsubo, Y., 2023. Collective Effervescence in Retail Investors: an Exploratory Inquiry to Understand the Meme Stock Bubble. Available at SSRN. https://ssrn.com/abstract=4321183.

Long, S., Lucey, B.M., Xie, Y., Yarovaya, L, 2023. I Just Like the Stock": the Role of Reddit Sentiment in the GameStop Share Rally. The Financial Review 58 (1), 19–37.

Pedersen, L.H., 2022. Game on: social networks and markets. J. financ. econ. 46, 1097–1119, 3.

Wu, Y., Liu, L., Shen, S., 2024. Did subsidiary's participation in paycheck protection program affect public parent company? Evidence from short selling. Econ. Lett. 241, 111791.