Social Media Dynamics of Shorted Companies

Carl Terve, Mattias Erlingsson, Alireza Mohammadinodooshan and Niklas Carlsson
Linköping University, Sweden
Emails: firstname.lastname@liu.se

***Abstract*—The discussions on social-media forums can impact the sentiment of a company, and consequently also its stock price. As we show here, some of the most shorted companies have provided some of the clearest examples of this relationship. In light of these observations, this paper presents a longitudinal study of the cross-forum dynamics of ten highly shorted stocks that saw significant discussions on the popular forums Reddit, Twitter, and Seeking Alpha. Using the posts from these forums, their sentiments, and the daily snapshots of the stock price of each company, we use a combination of qualitative case studies and quantitative hypothesis testing to derive new insights. Through a combination of time-series analysis, clustering, and domain-optimized sentiment analysis, we study the relationship between the times that discussions peak on the different forums, the changes in sentiment, and the stock price movements. We find that all three forums are likely to experience peaks in their activity close to each other, that Reddit is most likely to peak first, and that the sentiment of Twitter discussions were more sensitive to the current derivative of the stock price than the sentiment observed on the other forums.**

***Index Terms*—Shorted companies, Longitudinal analysis, Social media dynamics, Sentiment analysis**

I. Introduction

At the start of 2021, the world saw people inspired by the discussions on a Reddit channel taking on several big Wall Street investors [1]. At the time, several investors had heavily been *short selling* the GameStop (GME) stock, meaning that stocks were borrowed by investors and sold to someone else in the hope of later returning the debt at a lower price.[1] With many loans soon to expire, people on the Reddit channel `r/WallStreetBets` saw this as an opportunity to force these investors to lose significant money (and potentially make money themselves) by simply investing in the stock and pushing up the stock price. This movement took on its own life and the price of the stock skyrocketed (18-time increase from $17.25 to $325.00 between Jan. 4 to Jan. 29).

This is just one example where social media has played a big factor in the stock trading dynamics. There are also many examples of popular influencers such as Elon Musk [2], [3] or (pseudo anonymous) people claiming inside knowledge on these forums [4] whose posts have been found to substantially influence individual stock prices. It is therefore not surprising that investigations led by the U.S. Securities and Exchange Commission (SEC) in 2017 uncovered multiple cases where public companies hired communication firms and individuals to promote their stocks on the investment site Seeking Alpha.

[1]The investors make a profit from the price drop but lose money if they are forced to buy back the stock at a higher price or pay high interest on the loan if they do not give back the stock on time.

With more and more people engaging in discussions over these forums and the social interactions and emotions playing central roles in financial decision making [5], it becomes increasingly important to understand the relationship between these social media discussions and the stock trading dynamics. As exemplified by the GME case above, some of the most discussed companies are the short selling ones. One reason for this is that short selling primarily is done on companies that some investors find overvalued and that these investors often have released public reports explaining why they value the company much lower than the current market value, in the hope that this will help drive down the stock price.

James Chanos, for example, achieved significant profits in 2001 by combining short selling of Enron with aggressive criticism of their accounting procedures. Since then, several investors have effectively paired short selling with short reports explaining why a company is defective or overvalued. Andrew Left and his company Citron Research [6] are among the most successful and well-known short sellers, having published short reports on companies for over 20 years. Because of their historical performance, these investors' short reports could impact both short- and long-term stock prices.

The media flocked to the GME story because the Reddit community pushed back and ended up buying GME stock, forcing Citron Research and others to sell at loss. After closing its short position in GME, Citron Research said (via Twitter) that it will "no longer issue short reports" and instead focus on long holdings [7].

In this paper, we present the first cross-forum analysis of the social network dynamics of shorted companies. In particular, we present a longitudinal study of the cross-forum dynamics of ten shorted stocks that saw significant discussions on all three forums considered. The studied companies include the two companies that both have generated the most discussions on social media and that have been the most shorted companies in recent years (TSLA, GME), as well as the eight companies that Citron Research has released short reports about (GPRO, LYFT, MSI, NFLX, NVDA, SHOP, SNAP, TWTR) that also are of significant size and have seen the most discussions on Reddit and other social forums. (Table I, explained and discussed later, provides full names of the selected companies.)

We analyzed all posts about the above companies on Reddit, Twitter, and Seeking Alpha over the last 12 years (2009-2021), as well as daily stock statistics (e.g., volume and price). We used time-series analysis, clustering, and domain-optimized sentiment analysis to gain new insights from this data. For the sentiment analysis, each post was labeled using a custom

The definitive version was published in *Proc. International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Milan, Italy, Nov/Dec. 2022. https://doi.org/tbd.

TABLE I

DATASET SUMMARY BROKEN DOWN PER COMPANY OF INTEREST. (HERE WE SHOW NUMBER OF ORIGIN POSTS FOR EACH FORUM. FOR THE ANALYSIS WE ALSO INCLUDE COMMENTS ASSOCIATED WITH EACH ORIGIN POSTS. SEE FIG. 4 FOR THE INCREASES IN DAILY OBSERVATIONS.)

Ticker	Name	Country	Sector	IPO Year	Market Cap	Volume	Short Interest (%)	SA	Reddit	Twitter
TWTR	Twitter Inc.	US	Technology	2013	42 B	21,316,354	3.92	1,048	14,843	1,139,749
NFLX	Netflix Inc.	US	Consumer Services	2002	220 B	2,703,068	3.95	2,331	15,228	1,147,745
SHOP	Shopify Inc. Class A	CA	Technology	2015	136 B	1,085,378	3.43	241	2,815	203,903
LYFT	Lyft Inc. Class A	US	Technology	2019	16 B	4,210,679	10.71	89	3,225	74,774
MSI	Motorola Solutions Inc.	US	Technology	1977	34 B	387,647	2.23	52	181	34,504
SNAP	Snap Inc. Class A	US	Technology	2017	82 B	14,388,944	7.92	457	6,600	284,939
GPRO	GoPro Inc. Class A	US	Miscellaneous	2014	1.5 B	2,485,413	19.68	437	2,216	228,769
NVDA	NVIDIA Corporation	US	Technology	1999	353 B	5,414,764	1.71	1,038	8,927	619,243
GME	GameStop Corporation	US	Consumer Services	2002	13 B	7,407,663	167.30	146	206,071	1,062,097
TSLA	Tesla Inc.	US	Capital Goods	2010	556 B	32,289,304	16.17	810	42,180	3,485,851

sentiment-labeling method based on a state-of-the art model that considers the financial nature of the posts.

Our data-driven insights combine qualitative case studies and quantitative hypothesis testing. We apply quantitative hypothesis testing to provide support for and further insights into the identified cross-forum dynamics. Here, we focus on the relationship between the times that discussions peak on the different forums and the stock price movements.

Using time-series analysis and event clustering, we identify correlations between discussion patterns across forums. This includes patterns that reveal which forums see similar peaks and valleys in a company's discussions, and in what order. For example, we find that all three forums have peaks in activity that are closely related, with Reddit being the most likely to peak first. Compared to the other forums, Twitter discussions were more sensitive to the current derivative of the stock price.

Outline: The next two sections describe our company selection and data collection (Sections II-III). Next, the selected forums and companies are compared (Section IV) and our time-line analysis is presented (Section V). Finally, we present the related work and the conclusions (Sections VI-VII).

II. COMPANY SELECTION

To study the social-media dynamics of companies with shorted stocks, we identified stocks that in recent years (2019-2021) have been both highly shorted and seen significant discussions on social media. We then tracked both the stock price and social media activity about these companies backwards in time; either to around the time of their initial public offering (IPO) or to 2009, whichever happened most recently.

To limit our scope and simplify the analysis, we focus on the companies that (1) were listed on NYSE or Nasdaq, (2) had a market capitalization over $1B$ dollar, and (3) had a significant number of social media mentions during the period. (To obtain the social media posts, we used the methodology discussed in Section III.) From this subset we then selected ten significantly shorted companies. These were either hand-picked (GME, TSLA) or Citron Research had published short reports on them.

A. Social media discussion of shorted companies

GameStop (GME): Of the companies that meet our primary selection criteria, GME (as perhaps the most well-known shorted one) was the company for which we observed

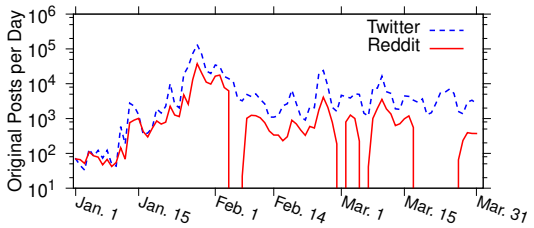


Fig. 1. GME (GameStop) posts on Twitter and Reddit during 2021-Q1.

the highest average short interest (167%), calculated as the fraction of loaned shares of the float, over the 2019-2021 period. Furthermore, as discussed in the introduction, the short selling of this stock was heavily discussed on social media at the start of 2021. This is seen in Fig. 1, where we can see that the discussion of GME increased by almost a factor of 1,000 times during January 2021, as the discussion of the short squeeze peaked towards the end of January. In fact, the short squeeze itself was largely engineered by Reddit users on r/wallstreetbets (as well as some hedge funds) decided to buy the GME stock so as to drive up the stock price and force short sellers, including Citron Research, to take significant losses.

Tesla (TSLA): Due to its high promise but lack of early revenue, one of the most shorted [8] and discussed companies in the world has been Tesla (TSLA). The company is also notable due to its CEO and co-founder, Elon Musk. With more than 100 million Twitter followers, his comments on Twitter have been found to have significant impact on both his own and other's companies stock prices [9].

Companies with short reports by Citron Research: Finally, we picked the eight companies that satisfied our selection criteria and had seen short reports by Citron Research [6].

Citron Research and other short sellers publish public *short reports* that explain and justify their short positions and/or company valuations. For example, Citron Research has published over 150 reports since 2001, with over 50 companies having seen regulatory interventions after the reports [6]. They have also influenced other people's valuations of the companies they report on, as well as discussions about these companies. For example, the public's reaction to a 2018 report about Twitter led to an 11% price drop on the same day [10].

For our selection, we focused on the Citron Research short reports published after 2009. Again, we limited our scope to companies that satisfied the three above mentioned criteria. To select a threshold for the minimum social activity, we used the number of original Reddit posts (not including comments).

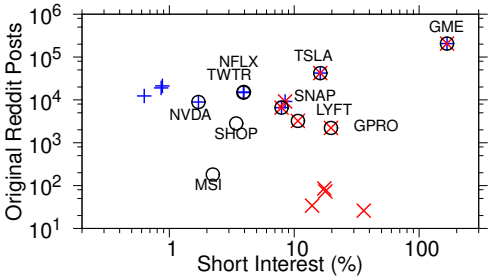


Fig. 2. Company comparisons using the total number of original Reddit posts and the average short interest (2019-2021). Here, the top-10 on social media are shown using blue plus signs (+), top-10 in terms of short interest are shown using red crosses (×), and the selected companies are shown using black circles (○) with individual ticker labels. (NFLX and TWTR have overlapping markers.)

Ranking the companies based on this metric, we identified a big gap in the activity level observed for different companies and set the threshold to be within this gap. This resulted in eight additional companies being selected.

B. Perspective: Comparing social activity and short interests of selected companies

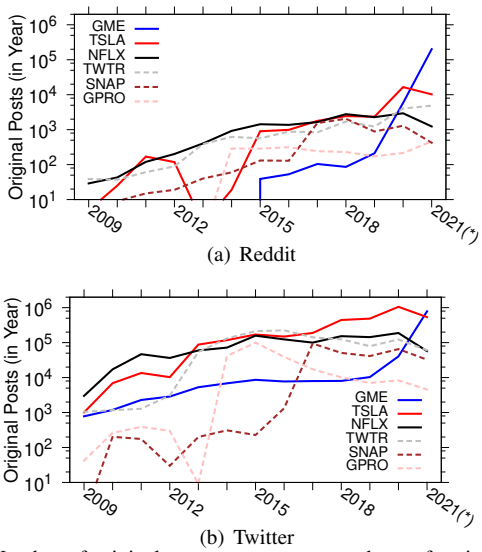
To put the shorting and social network activity of the selected companies into perspective, we identified (1) the sets of stocks that were most observed on Reddit's stock related channels between 2019-2021 and (2) the set of stocks that saw the highest average short volume ratio over the same time period. We again restricted ourselves to the above three mentioned criteria. We also filtered away companies with an IPO younger than 2019.

Most discussed companies: For social media activity, we again use the number of original Reddit posts about each company (see Section III for collection methodology) and limit the collection to the period 2019-01-01 to 2021-04-20.

Most shorted companies: To determine the most shorted companies, we calculated the average daily short sale volume and total trading volumes using data collected from Financial Industry Regulatory Authority (FINRA) [11]. Here, the daily short volume represents all reported shares being sold short together with short sales exempt trades [12]. While the peak volumes of a company can be considerably higher than its average values, the average metric provides a good view of the (on average) most shorted companies over the last two years and helps us put the selected companies in some context.

High-level comparison: Fig. 2 shows the top-10 companies ranked based on social activity (blue + markers) and short interest (red × markers) together with our company selection (black ○ markers). Note that our company selection includes the two most discussed companies GME and TSLA, and that many of the other companies have both relativity high social activity and (average) short interest. In fact, none of the other companies included in the plot (i.e., from the top-10 most discussed companies or the top-10 most shorted companies) scores higher than any of the 10 selected companies along both dimensions.

Historical lookback: To help explain why the problems discussed here have not received significant attention until



(a) Reddit

(b) Twitter

Fig. 3. Number of original posts per company and year for six companies. ((*) 2021 only includes data for 4 months.)

now, we calculated the corresponding statistics for the top-10 companies with regards to Reddit posts and seven most shorted companies a decade before our primary analysis period (i.e., 2009-01-01 to 2011-04-20). Due to not being able to find the short volumes for all companies during this time period, we restricted this analysis to the most shorted stocks on the S&P 500. While this gives us a somewhat smaller pool of companies, the comparison was striking as only a single company had more than 350 Reddit posts over the period (Ameriprise Financial (AMP) with 1,339 posts) and only one of the seven shorted companies was observed in more than seven posts (American International Group (AIG) with 326 posts, which saw a stock price drop of 97% during 2007-2008 just before our period of interest). Of the companies studied in this paper, the companies with the most original Reddit posts over the 2009-2011 period were TWTR (39), NFLX (29), and NVDA (24). All others had less than ten original posts over that time.

Although the total number of origin posts on Twitter was higher 10 years earlier, also Twitter has seen a significant increase in stock discussions. To illustrate the rapid increase in the discussions over the last 10+ years, Fig. 3 shows the number of origin posts per company and year for six example companies. First, we observe substantial increases in the total number of origin posts per year. (Y-axis on log scale.) This again highlights the quickly increasing impact that social media can have on stocks. Second, the Twitter data (which has the highest volumes) shows clear periods of increased activity for: GPRO (2015), SNAP (2017), and GME (2021). Citron Research took short positions of each of these companies around these times. It can also be noted that the trends for TWTR, TSLA, and NFLX are generally increasing.

Besides the increased volumes showing that the importance of this work is increasing, the increased volumes also enable analysis that might not have been possible previously. As an example, the limited social media activity 10+ years ago,

prevented us (and others) from looking more closely at historical social media dynamics experienced by these companies. Therefore, for the most part, we report stats for the more recent periods 2016-2021 or 2019-2021.

III. Data collection and per-post labeling

In addition to collecting all Reddit, Twitter, and Seeking Alpha posts about the companies of interest, we also collected daily stock metrics (e.g., open, high, low, close, volume, and adjusted volume) about each company. Furthermore, we used a state-of-the-art sentiment analysis model to assign a sentiment value to each post. We next motivate our forum selection and explain how the data collection and labeling were performed.

A. Data collection

Social media selection: Reddit is one of the most popular social media domains (Tranco rank of 36 [13] as of 2022-06-17) and has seen much media attention due to its central role in influencing the stock price of GameStop (GME) [14]. Twitter is highly popular (Tranco rank of 9) and has a similar user base (e.g., as calculated from analysis of familiar visitors and search keywords [15]). In contrast to the first two forums, Seeking Alpha – which is highly popular among stock investors – is much more domain specific and is dominated by professional posters. Seeking Alpha has 20 million monthly users, 7,000 monthly contributors, publishes around 10,000 investing ideas per month, each reviewed before publication [16]. Compared to posts on Reddit and Twitter, these reports are usually longer, directly comment on a stock, and target other investors.

Reddit collection: We used the API wrapper PMAW [17] to query the Pushshift database for all historic Reddit posts associated with each company of interest. Specifically, we used a query based on the company's stock symbol and company name for the period 2009-01-01 to 2021-04-20. To reduce pollution, we only queried 17 carefully selected subreddits related to the keywords *finance*, *stocks*, and *investing*. The choice of subreddits was based on the keyword searches using Reddit's search functionality, which maps to the most relevant or most commented subreddits. Selected subreddits include *r/investing*, *r/wallstreetbets*, and *r/stocks*.

Twitter collection: We used the Twitter full-archive API endpoint [18] using the cashtag functionality together with the ticker of each company. This approach provided an effective way to fetch tweets relevant to finance, stocks, and investing. To limit the scope and avoid data overlap, we only retrieved English-speaking tweets and excluded retweets.

Seeking Alpha collection: For Seeking Alpha, we used the Rapidapi.com API [19]. In particular, we used the endpoints `analysis/list` and `analysis/get-details`.

Daily stock values by Morningstar: For each company of interest, we collected Morningstar's recordings of the daily values for open, high, low, close, volume, and adjusted volume. This data was also complemented with company fundamentals such as Short Interest as a percentage of shares float, Price/Cash, and Price/Sales. The 10-day volatility for each company was also included. These parameters help analyze a company's financial and economic position.

B. Sentiment labeling

Pre-processing: For the most part, the pre-processing was similar among the platforms. First, we unscaped the HTML special characters, changed to lower case letters, and replaced the contractions (e.g., "ima") with their original form (e.g., "I am going to"). Then we employed platform-specific regular expressions for locating and replacing usernames, numbers, URLs, and cashtags with related constants. For example, all cashtags were replaced with the constant "company". For hashtags, however, we only removed the hashtag sign but kept the original text. Furthermore, for all the platforms, regular expressions were used to detect most cases of the happy (e.g., `U+1F600`) and sad (e.g., `U+1F614`) emojis and replace them with the constants "happy" and "sad". The same procedure was followed for the sadness and happiness related emoticons.

Sentiment extraction: For sentiment analysis, we used FinBERT [20]. This model takes advantage of the very good performances that has been demonstrated by recent pre-trained transformer-based language models (e.g., BERT) for NLP tasks but is further adapted for the financial domain. Building on BERT, it is further trained for the financial domain using TRC2 data [21] Furthermore, for the downstream task of financial sentiment analysis, FinBERT is finally finetuned on the Financial PhraseBank dataset [22]. The final output of the model gives the magnitude of the neutral, negative, and positive sentiments for the input text (the magnitude range is 0.0 to 1.0). We consider the sentiment with the most significant magnitude when classifying the input. For our analysis, we calculated a combined value over all sentences based on the number of positive, negative, and neutral sentences.

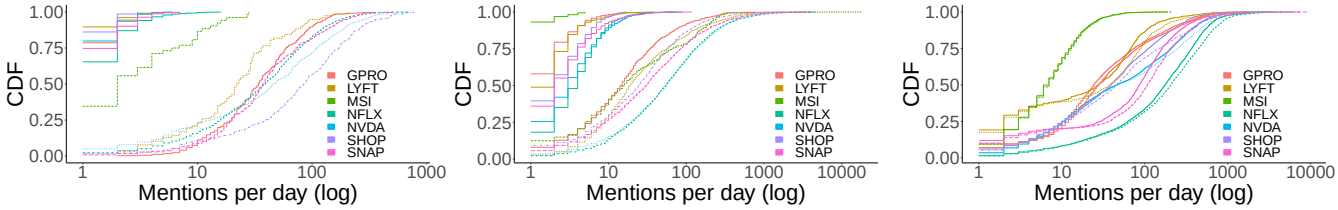
IV. Company and forum comparisons

A. Dataset summary

In total, we collected and analyzed all posts and comments associated with 302,286 original Reddit posts, 6,649 original Seeking Alpha posts, and 8,281,574 original Twitter posts. Table I summarizes the collection statistics about each target company. In addition to the number of social network posts about each company (last three columns), the table also shows each company's ticker, the country it is registered in, the sector it belongs to, its IPO year, market cap, trading volume, and average short interest. We note that among the three companies with the relatively highest trading volume compared to their cap (GPRO, GME, TWTR), the short interest is also highest for two of these companies (GPRO, GME).

B. Forum comparison

While Seeking Alpha's origin posts are larger and more infrequent, many of them attract significant forum participation. Twitter posts, on the other hand, are typically short and receive fewer interactions. Reddit is a middle ground. For each forum, Fig. 4 shows the number of company mentions in original posts (solid curves to the left) and comments (dotted lines to the right). While the number of original posts varies between forums (as shown in Table I), the volume varies less when comments are included. Moreover, forum activity is significant



(a) Seeking Alpha (b) Reddit (c) Twitter

Fig. 4. Comparison of the number of company mentions in origin posts and comments for the three different forums.

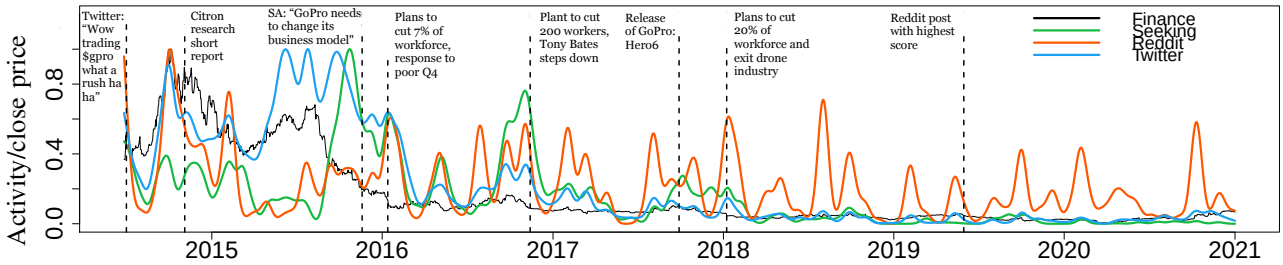


Fig. 5. GPRO time series with annotated external events, stock value, and smoothened social media volume.

for most companies and days. Yet, we deemed the volume of original posts to be sufficient for the longitudinal analysis (next section) of both Twitter and Reddit, and therefore only included comments for our analysis of Seeking Alpha.

V. Time-line analysis

For our time-line analysis, we *smooth* the data, identify *local extreme points*, and *cluster* the extreme points associated with different social media activity metrics and stock metrics.

A. Smoothed time series analysis

Kernel smoothing: Rather than looking at day-to-day fluctuations, we aim to capture underlying trends at larger time scales. For this purpose, we used kernel smoothing. This choice is motivated by the volatile nature of the data and the desire to not have to make any assumptions about the underlying probability distributions. With kernel smoothing, the regression function is estimated using kernel density estimation without introducing a parametric model [23]. For simplicity, we use the univariate kernel density estimator:

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\{\left(\frac{x - X_i}{h}\right)\}, \qquad (1)$$

where K is the smoothing kernel deciding how neighboring points are weighted, satisfying $\int K(x)dx = 1$, h is the smoothing bandwidth, and X_i are the random samples. After choosing a kernel K, h is tuned for desirable smoothness of the estimations. The calculations were done using the *ksmooth* package of R, using the default kernel (called *normal*, which itself uses the Gaussian density function) and $h = 14$ days.

Example timeline: Let's now consider the four primary time series of an example company. Fig. 5 shows these time series for GoPro Inc. (GPRO). Here, we include both a smoothed time series of the company's stock value (from when it was listed) and the social media activity for the platforms Twitter, Reddit, and Seeking Alpha. To simplify visualization, all values are normalized relative to their (global) peak values during the time period. We also annotated the figure with vertical lines indicating external events such as the release date of a short report by Citron Research on the company. In this report, they argued for a drop in price from $79 to below $30 within the next 12 months, something that indeed ended up happening. To address these and other concerns, the company decided to change its business model later that year.

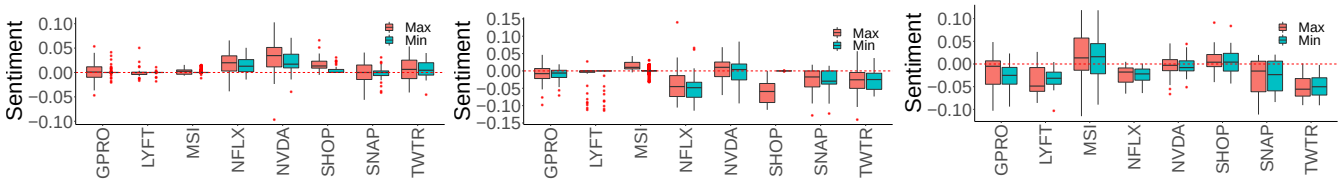
We note that the short report came at the time that the stock was at its all-time peak and at the backend of a peak in social network activity (although the post seems to have generated some renewed interest on Seeking Alpha). While the stock saw some up-and-downs during 2015, there was a clear down trend after the short report. In retrospect, and perhaps not surprisingly, the most commented post on Seeking Alpha was quite negative, stating that "GoPro needs to change its business model". This post took place on Nov. 19, 2015, when the stock already had seen a significant downfall, and already on Jan. 13, 2016, the company announced plans to cut 7% of its workforce due to poor fourth-quarter results. Other company announcements and financial reports can usually be linked to significant social activities (some annotated in Fig. 5).

To study the dynamics across the different forums and the stock prices, we next identify the extreme points of each time series and then analyze how they relate.

B. Peak and valley analysis

Identifying highs and lows: Given the smoothed curves, we next identify local highs and lows. In addition to representing peaks and valleys in the level of social media activity or stock price, these turning points represent trend changes. We used the *local.min.max()* function in the *spatialEco* package of R to identify these local minimums and maximums.

Sentiment-based comparisons: Fig. 6 compares the average sentiment across different forums during various extreme events. We comment on two observations. First, in most cases, the median sentiment associated with the peaks (max) have similar or slightly higher sentiment than the valleys (min).



(a) Seeking Alpha (b) Reddit (c) Twitter

Fig. 6. Comparison of the average sentiment as seen on the different forums at the time of different extreme events.

TABLE II
CLUSTER-BASED ANALYSIS. ONE-SIDES BINOMIAL HYPOTHESIS TESTS THAT THE ANY CLUSTER COMPOSITION IS EQUALLY LIKELY.

Reddit	S Alpha	Twitter	N	p_{bin}	z_{norm}	95% test (p=1/8)
max	max	max	85	$< 10^{-8}$	15.42	Confident **above**
min	min	min	44	$5.6 \cdot 10^{-7}$	5.08	Confident **above**
min	max	min	14	0.099	-1.31	Not rejected
max	min	max	7	$4.2 \cdot 10^{-4}$	-2.99	Confident below
max	max	min	4	$6.9 \cdot 10^{-6}$	-3.71	Confident below
min	min	max	2	$1.5 \cdot 10^{-7}$	-4.18	Confident below
max	min	min	1	10^{-8}	-4.42	Confident below
min	max	max	1	10^{-8}	-4.42	Confident below

The positive differences match the intuition that the sentiment may be higher during periods with more activity (i.e., peaks) than during periods with less activity (i.e., valleys) and the relatively small size in the differences captures that both types of extreme points capture periods in which the derivative of the interest in a stock is changing from positive to negative and vice versa. Second, only MSI has positive median sentiment for all three forums. In general, the sentiment differences between the forums are significant for most companies, and while it is not always the same forum that is the most positive/negative, Seeking Alpha tends to be less negative than Reddit and Twitter.

C. Clustering of extreme points

To study cross-platform dynamics, we next clustered local extreme points based on their relative time of occurrence. We first extract and place all detected extreme points from each time series (across the three forums) into a shared list including all extreme points linked with that company. Using this one-dimensional series of sorted dates, we then partition the data into k clusters using optimal k-means clustering. This was effectively implemented using dynamic programming [24], which minimizes the sum of squared distances from each cluster element to its cluster mean. (Here, we used the implementation provided by the R package *Ckmeans.1d.dp*.) After experimenting with different k values, we decided on a heuristic that sets k equal to 1.2 times the number of extreme points observed within the time series of a company's stock price after smoothing. The 20% inflation was used to increase the odds that each financial event was included in a separate cluster (together with its associated social media events). There are sometimes also local extremes from each time series that are not associated with any other extreme point. For these cases, the 20% inflation in clusters helps isolate such extremes.

For simplicity, we only analyze the clusters that have precisely one Twitter (T), one Seeking Alpha (S), and one Reddit (R) extreme point that all are non-zero. Table II summarizes the clustering of extremes observed 2016-2021. Perhaps the most noticeable result is that max points are highly clustered with other max points, and min points are highly clustered with other min points. For example, out of the 160 clusters, 85 clusters consist of only local maximums and 44 clusters consists of only minimums. The probability of this happening if the clusters were unbiased is very close to zero (z-scores of 15.42 and 5.62, respectively, if using binomial testing with normal approximation for large numbers). This shows that at a time that a stock sees an extreme level of activity in a forum it is likely to see an extreme level of activity (at least closeby in time) also on other forums.

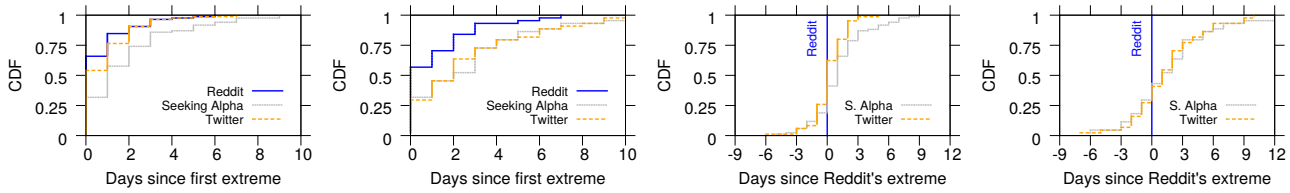
The observed bias is substantial. To put the bias in perspective, we note that finding this many clusters consisting of only maximum points would result in a rejected null hypothesis at the 95% confidence level even if the probability of such cluster would be $p = 0.46$. This is much greater than the unbiased probability of p=0.125 (tested for above).

The observed bias also appears invariant to time period considered and the bandwidth (h) used. For example, when instead considering the first six years (2009-2015), the "all-maximums" clusters and the "all-minimums" clusters are responsible for 39 and 25 out of the 82 clusters identified during this period (with h=14). The probability to observe such bias for a particular class if cluster composition was unbiased is close to zero ($< 10^{-8}$ and $1.4 \cdot 10^{-5}$, respectively, with z-scores of 9.43 and 4.76, respectively). Finally, when considering the bigger bandwidth of $h = 30$ over the full period, the corresponding values where 65 and 38 out of 116 clusters (both p-values $< 10^{-8}$, with z-scores of 14.04 and 6.46, respectively).

D. Order and timing of shared extreme points

The order that the extreme points within a cluster with shared extreme points ("all maximums" or "all minimums") is reached may provide insights into which forums may be leading or reacting to the discussions of certain companies.

Shared max-peak clusters: First, it should be noted that in 35 out of 85 max-peaks cluster cases, the first peak is shared between at least two forums, again highlighting the close tie in discussion activity between forums. Second, and among the set of cases with a single forum peaking first, Reddit is the most frequent *winner* (24 out of the remaining 50 cases) and Seeking Alpha is the least frequent *winner* (10 cases). The relative win frequencies look similar when considering the shared leading cases. In this case, Reddit is (shared) *winner* in 56 out of the 85 cases, Twitter in 46 cases, and Seeking Alpha in 27 cases. The bias toward Reddit being first to peak is



(a) Max points vs. first max (b) Min points vs first min (c) Max points vs max of Reddit (d) Min points vs min of Reddit

Fig. 7. CDFs of the time (in days) of a forums extreme point and the day of either the first extreme point in the cluster or the day that Reddit had its extreme value. (2016-2021 data with h=14.)

TABLE III

SENTIMENT COMPASSION OF UP-PERIODS AND DOWN-PERIODS. TO TEST THE BIAS THAT EACH FORUM IS RELATIVELY MORE LIKELY TO HAVE A POSITIVE SENTIMENT DURING UP-PERIODS THAN THEY ARE TO BE POSITIVE DURING A DOWN-PERIOD WE USE BINOMIAL HYPOTHESIS TESTS.

	Increasing price			Decreasing price			Test stats		
Forum	Above	Below	Samples	Above	Below	Samples	z_{obs}	p-value	95% test
Twitter	131	107	238	78	161	239	4.93	$4 \cdot 10^{-7}$	Confident **affected**
Reddit	129	109	238	117	122	239	1.147	0.12	H_0 not rejected
Seeking Alpha	103	135	238	79	160	239	2.30	0.011	Confident **affected**

significant. For example, as reference points, the probability of getting 24 or more out of 50 if there was no bias (i.e., p=1/3) is 0.022, and the observation of observing 56 (or more) out of 85 is rejected at 95% confidence up-to p=0.56.

Shared min-valley clusters: Shared min-valley clusters had a similar order. However, a shared day of the initial minimum point was rare (only 5 out of 44 cases). Reddit led 20 out of 39 of the remaining cases. Similarly, looking at the number of times each forum was first (including ties), Reddit was first in 24 out of the 44 cases, Twitter in 13 cases, and Seeking Alpha in 13 cases. Also here, the bias was significant at the 95% level (p=0.015) with 20 out of 39 cases.

Timing-based characterization: The above numbers suggest that max clusters have a high concentration of extreme values. This becomes clearer when looking at the CDFs of when each forum reaches its peak in a max cluster or its minimum in a min cluster. These distributions are shown in Figs. 7(a) and 7(b), respectively. Here, we also clearly see that Reddit is more likely to reach such extreme point ahead of the others. However, there are also several instances where the other forums are well ahead of Reddit. This is illustrated in Figs 7(c) and 7(d), which shows the timing of the other two forums extreme value point relative to that of Reddit's extreme value point.

E. Impact of stock movements

Unsurprisingly, we see more positive sentiment for stocks when the price rises than when it falls across all three social media forums. However, which forum's sentiment is most influenced by up-trends? To answer this question, each forum was individually tested with binomial hypothesis tests. Here, we tested the null hypothesis that a time period with increasing stock value is equally likely to have above-average sentiment as a time period with decreasing stock value. Table III shows the results. While the null hypothesis (as expected) is rejected for both Twitter and Seeking Alpha, it is not rejected for Reddit. This suggests that the sentiment on Reddit is much less sensitive to fluctuations in the stock prices than the other two forums, for which the sentiment indeed is affected by the changes in stock prices. Looking at the p-values we also see that Twitter is the forum where the sentiment is most affected by the current stock movement, and Reddit the least.

We did not observe any significant correlations between the sentiment and the up-trends and down-trends in social forum activity. This confirms the intuition that the derivative of stock price has a greater impact on current forum sentiment than the derivative of stock discussion level.

VI. RELATED WORK

Financial prediction using social media: Much research has demonstrated how social media can be used for market predictions. While this is not our goal, we note that related work with this goal has mainly targeted predication of two market variables: the return (price) and the volatility of the stocks. As an example from the first category, Chen et al. [25] show how sentiment data on Seeking Alpha can predict future stock returns. Focusing on Twitter, Tan et al. [26] demonstrate that the positive tone of posts has greater predictive power in small and developing market companies than in large ones. Using r/WallStreetBets, Huynh et al. [27] demonstrate that improved price forecasting can be achieved through careful trust screening of users and records, and by optimizing the time window used for prediction. Telli et al. [28] use Reddit and Wikipedia data to examine how public interest affects the cryptocurrency markets.

To study both price and volatility, Pedersen [29] presents a platform-independent, theoretic model. Here, three groups of investors are considered: (1) naive investors who learn via a social network, (2) "fanatics" (potentially retail investors) spreading fake news, and (3) rational short- and long-term investors. Using this model, it is then demonstrated that the interaction of the investor groups can directly impact future share price movements and generate extra volatilities.

Another example of a recent stock volatility study is the work by Frino et al. [30], who find positive correlations between volatility and social media posting activity and divergence of opinions as observed on Twitter. Jiao et al. [31] also highlight that strong social media coverage predicts high subsequent return, volatility, and trading activity, but that high news media coverage predicts the opposite. Tafti et al. [32]

show that Twitter also is able to adequately provide a broad and global livestream of market information.

Shorting and social media: Few research studies have examined the influence of social media on the dynamics of shorted stocks. Among them, Hu et al. [33] explore the impact of several dimensions of social media activity on stock-related factors, including short selling. The authors show that higher Reddit traffic, positive tone, and the connectedness of the posts indicate reduced shorting flows the following day. There is also additional research aimed at providing deeper knowledge into the recent GameStop situation. For example, Anand and Pathak [34] report that there was a statistically significant impact of Reddit sentiment 10 minutes in advance on the GME volatility. In terms of price dynamics, Umar et al. [35] show that the Reddit sentiment may have positively contributed to the GameStop returns, highlighting the potential short return in following these social media groups. Switching focus from the U.S. market to the Chinese stock market, Cui et al. [36] demonstrate that social media sentiment is unusually optimistic prior to heavy short interest. Once highly shorted, the sentiment becomes excessively negative.

In contrast to prior work, we study and provide insights into the relative dynamics between the different forums when it comes to individual shorted stocks.

VII. Conclusion

We have presented a longitudinal study of the cross-forum dynamics of ten shorted stocks. Our characterization captures the dynamics between discussions and stock prices. Here, we employ a combination of time-series analysis, clustering, and domain-optimized sentiment analysis to investigate the relationship between the times of discussion peaks on various forums, sentiment changes, and stock price movements. Our findings include the following: all three forums examined here (Reddit, Twitter, Seeking Alpha) are likely to experience closeby peaks in activity, Reddit is most likely to peak first, and Twitter sentiment is more sensitive to the current stock price derivative than that of other forums. Our findings highlight differences between forums and shed light on cross-forum trends that may aid in the prevention of market manipulation.

References

[1] B. Brumberg, "Reddit and GameStop lessons: Former SEC enforcement chief explains stock manipulation and how to avoid trouble," *Forbes Magazine*, Feb. 2021.

[2] L. Hooker, "Tesla share price falls after Elon Musk's Twitter poll," https://www.bbc.com/news/business-59209942, Nov. 2021.

[3] S. Shead, "Elon Musk's tweets are moving markets — and some investors are worried," https://www.cnbc.com/2021/01/29/elon-musks-tweets-are-moving-markets.html, Jan. 2021.

[4] SEC, "Updated investor alert: Social media and investing – stock rumors," https://www.sec.gov/oiea/investor-alerts-bulletins/ia_rumors.html, Nov. 2015, accessed: Oct. 2022.

[5] J. R. Nofsinger, "Social mood and financial economics," *Jrn. Behav. Financ.*, vol. 6, no. 3, pp. 144–160, 2005.

[6] A. Left, "About citron research," https://citronresearch.com/who-is-citron-research/, 2017, accessed: Oct. 2022.

[7] M. Fox, "Citron Research says it will stop publishing short-seller research after the GameStop squeeze," https://markets.businessinsider.com/news/stocks/citron-stop-publishing-short-seller-research-gamestop-squeeze-andrew-left-2021-1, Jan. 2021.

[8] M. J. Coren, "Even after GameStop, Tesla remains the most shorted stock in the world," https://qz.com/1979325/tesla-not-gamestop-is-the-most-shorted-stock-in-the-world, Mar. 2021.

[9] N. Dailey, "Elon Musk is the person of the year largely because of his effect on the world of finance," https://markets.businessinsider.com/news/stocks/elon-musk-top-12-tweets-moved-markets-tesla-stock-doegcoin-2021-12, Dec. 2021.

[10] L. Feiner, "Twitter plummets after citron research calls company 'toxic' following amnesty international report," https://www.cnbc.com/2018/12/20/twitter-stock-down-after-citron-research-calls-it-toxic.html, Nov. 2018.

[11] FINRA, "Short sale volume daily," https://www.finra.org/finra-data/short-sale-volume-daily, 2021, accessed: June. 2021.

[12] ——, "Short sale volume daily data user guide," https://www.finra.org/sites/default/files/2020-12/short-sale-volume-user-guide.pdf, 2021, accessed: May 2021.

[13] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proc. NDSS*, 2019.

[14] K. Grant, "GameStop: What is it and why is it trending?" https://www.bbc.com/news/newsbeat-55841719, Jan. 2021.

[15] Alexa, "reddit.com competitive analysis, marketing mix and traffic," https://www.alexa.com/siteinfo/reddit.com, 2021, accessed: Mar. 2022.

[16] Seeking-Alpha, "About Seeking Alpha," https://seekingalpha.com/page/about_us, accessed: Feb. 2022.

[17] M. Podolak, "PMAW: Pushshift Multithread API Wrapper," https://github.com/mattpodolak/pmaw, accessed: Jan. 2022.

[18] Twitter, "Search API: Premium," https://developer.twitter.com/en/docs/twitter-api/premium/search-api, accessed: Jan. 2022.

[19] A. Dojo, "Seeking alpha API documentation free with API key & SDK," https://rapidapi.com/apidojo/api/seeking-alpha, 2021, accessed: May 2021.

[20] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv:1908.10063*, 2019.

[21] NIST, "Reuters corpora (rcv1, rcv2, trc2)," http://trec.nist.gov/data/reuters/reuters.html, accessed: Oct. 2022.

[22] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. for Inf. Sci. and Tech.*, vol. 65, no. 4, pp. 782–796, 2014.

[23] M. Wand and M. Jones, *Kernel Smoothing*. Chapman & Hall, 1995.

[24] H. Wang and M. Song, "Ckmeans.1d.Dp: Optimal k-means clustering in one dimension by dynamic programming," *The R journal*, vol. 3, no. 2, pp. 29–33, 2011.

[25] H. Chen, P. De, Y. j. Hu, and B.-H. Hwang, "Wisdom of crowds: The value of stock opinions transmitted through social media," *Rev. Financ. Stud.*, vol. 27, no. 5, pp. 1367–1403, 2014.

[26] S. Duz Tan and O. Tas, "Social media sentiment in international stock returns and trading activity," *Jrn. of Behav. Finance*, vol. 22, no. 2, 2021.

[27] D. Huynh, G. Audet, N. Alabi, and Y. Tian, "Stock price prediction leveraging reddit: The role of trust filter and sliding window," in *Proc. IEEE Big Data*, 2021.

[28] Ş. Telli and H. Chen, "Multifractal behavior relationship between crypto markets and wikipedia-reddit online platforms," *Chaos, Solitons & Fractals*, vol. 152, p. 111331, 2021.

[29] L. H. Pedersen, "Game on: Social networks and markets," *Jrn. of Finan. Economics*, 2022.

[30] A. Frino, C. Xu, and Z. I. Zhou, "Are option traders more informed than twitter users? a pvar analysis," *Journal of Futures Markets*, 2022.

[31] P. Jiao, A. Veiga, and A. Walther, "Social media, news media and the stock market," *Jrn. of Econ. Behav. Organ.*, vol. 176, pp. 63–90, 2020.

[32] A. Tafti, R. Zotti, and W. Jank, "Real-time diffusion of information on twitter and the financial markets," *PloS one*, vol. 11, no. 8, 2016.

[33] D. Hu, C. M. Jones, V. Zhang, and X. Zhang, "The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery," *Available at SSRN 3807655*, 2021.

[34] A. Anand and J. Pathak, "The role of reddit in the gamestop short squeeze," *Economics Letters*, vol. 211, p. 110249, 2022.

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

[36] G. Cai, R. D. McLean, T. Zhang, and M. Zhao, "Short sellers in the realm of social media: Arbitrageurs or manipulators?" *Available at SSRN 3907480*, 2021.