Social Stock Market Cascades

Nitzan Cohen, Rina Galperin

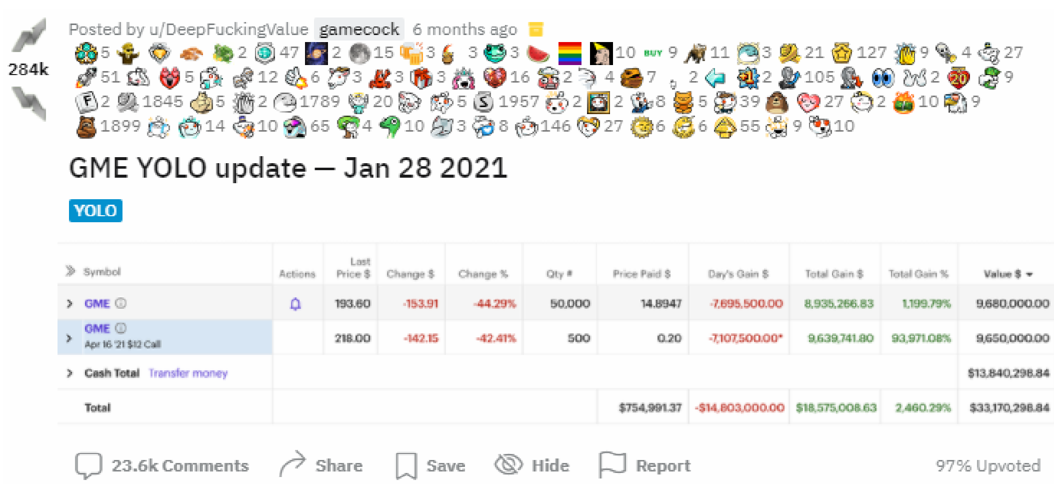
July 2021

1 Introduction

GameStop is an American video game, consumer electronics, and gaming merchandise retailer located in the US and a public company traded in the stock market (*i.e.,* GME stock) [1]. Hedge funds are a group of investors that are supervised by a money manager. These hedge funds aim to make profits by short selling on failing stocks. A big hedge fund has a pool of investment money that gives the investor the bandwidth to invest aggressively and make complex investments for bigger payouts. Over the years, GameStop's stock had been falling.

Reddit is a social media platform founded in June 2005. Like many other social media websites, contributors post content and other users can add comments in response to the original post. The Reddit community is a collection of forums, where each forum is dedicated to a particular topic called a subreddit. **Wallstreetbets (WSB)** [7] is a very popular subreddit that was created on January 31, 2012 with a particular emphasis on highly speculative trading strategies.

In January 2021, Reddit users noticed hedge funds were heavily short selling the stock, particularly the $13 billion hedge fund Melvin Capital [2]. A certain Reddit thread (Figure 1) was posted as part of an act of protest against the hedge funds. The social event caught the attention of many young online traders, which led to the share prices increasing to unwarranted levels, causing a cascade in the social network and reaching an uncharacteristic number of people. In order to attempt and stabilize the market, some brokers stopped the option to trade in the GameStop stock. As a result, the event snowballed, increasing the interest in the stock even more.



284k

Posted by u/DeepFuckingValue gamecock 6 months ago

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

23.6k Comments Share Save Hide Report 97% Upvoted

Figure 1: The WSB post that triggered the GME incident

2 Research Scope

Research Questions: Following the GameStop incident we were motivated to understand how certain events in the social context affect stock values and what are the main characteristics of such events.
Contribution: We provide a methodology for mapping trends across a social network, thus defining the boundaries of a social network event (start and end times of the event). We measure the correlation between our self-defined metrics and the stock's price. Moreover, we show an analysis of the several stocks (AG, GME) in specific timeframes as proof-of-concept.

3 Previous Work

Social media generates plenty of real-time content at a large rate. From all the content that people create and share, only a few topics manage to attract enough attention to rise to the top and become be classified as *trends*. In [9], the authors conducted an intensive study of trending topics on Twitter and provided a theoretical basis for the formation, persistence and decay of trends. They demonstrated empirically how factors such as user activity and number of followers do not contribute strongly to trend creation and its propagation. In fact, the authors found that the resonance of the content with the users of the social network plays a major role in causing trends. As a result, in our work we chose to not factor user-related properties but rather focus on the attributes of the post and its content (Section 5.1). In [12] the authors focused on long trends, *i.e.,* why some of the topics discussed in social media stay popular for longer periods of time than others and thus contribute to the social agenda. The authors provided a threshold criterion that predicts the long term persistence of social trends. The predictions of the stochastic dynamical model were compared with measurements taken from Twitter. The authors defined the cumulative count of tweets and re-tweets that are related to a certain topic as a proxy for the popularity of the topic. We chose to use a similar definition for the importance of the content, using the score of a comment (upvotes minus downvotes) (Section 4). The idea of social media begin able to influence real-life fields and events is not a new concept. Extensive research has been done on the impact of social media on our lives in general and on the stock market specifically. [11] discussed how emotional sentiment plays an important role in investment decision making. Emotional sentiment about a firm's stock that spreads rapidly through social media is more likely to be incorporated quickly into stock prices. The authors analyzed the cumulative sentiment (positive and negative) in 2.5 million Twitter posts about individual $S\&P500$ firms and compared this to the stock returns of those firms. They concluded that there was a significant impact on the stock's returns on the next trading day, the next 10 days, and the next 20 days. Importantly, the authors claim that this holds regardless of the user's popularity (*i.e.,*, number of followers), thus supporting both the conclusions of [9] and our chosen approach. In our work we also explored the sentiment of the comment and whether or not it has a connection to the stock-related action contained in the comment (Section 5.2.1). A recent work was also inspired by the GameStop incident and addressed the market consequences of investment advice on Reddit's WSB [10]. The authors showed findings that suggest that recommendations on WSB contain significant investment value, despite the conventional view that the platform primarily attracts uninformed investors. They find that due diligence reports – reports that are vetted by moderators as pertaining some analysis – contain investment value. With this conclusion in mind, we also look into the impact of WSB on the stock market and demonstrate our approach on several stocks in specific time-frames as proof-of-concept (Section 5).

4 Our Approach

We want to estimate how valuable a discussion was to the exposure of a given topic. We define a local popularity metric (Section 4.1.1) to extract all the topics within a post, taking into account both the frequency of mentions and the importance of the information contained in the comments (estimated by the score). When examining the posts across the network, we can rank the stocks that will serve as candidates for the main "hot topics". Such hierarchy on its own does not suffice to conclude an existence of a trend.

Local popularity does not address the global magnitude of a post compared to other posts. Therefore we define the exposure metric (Section 4.1.2), used as a means of comparison between two different posts as a function of the stock symbol. Since exposure is a function of time, we can define the *stock market social network trend* using various strategies over a graph $Exposure(symbol = s, time)$, for example by looking at peaks of the exposure. Given a stock s, we can compute $Exposure(s)$ over a period of time and the correlation (measured by non-linear Spearman Correlation) with the price of the stock in that period of time. The correlation can be further used to understand the impact of a certain trend.

4.1 Metrics

4.1.1 Local Popularity

Defines the popularity of a symbol relative to other symbols within the post. *I.e.,* we look at the distribution of stock symbols in the post to extract the main topic.

Estimates the probability that a given post is about a certain stock symbol, where the comments are extracted up to a given date.

$$LP'(symbol, post) = \sum_{height=1}^{n} 2^{-height} \sum_{comment \in level} \mathbb{1}[symbol \in comment] * S(comment)$$

where $S(comment)$ is the number of upvotes minus downvotes for the comment.
Since we want to compute a hierarchy of topics within the discussion, we impose:

$$\sum_{symbol \in post} LP'(symbol, post) = 1$$

The overall vector of topics distribution is defined as follows:

$$LP(post) = softmax([LP'(s_1, post), ..., LP'(s_m, post)])$$

4.1.2 Exposure

Defines the contribution of a given thread to a specific stock's exposure in the social network, in a given day.

$$Exposure'(symbol, post, day) = NC(post, symbol) * LP(symbol)$$

where NC is the number of new comments posted about the stock in the given day.
The cumulative contribution of all the active posts in the given day is defined as follows:

$$Exposure(symbol, day) = \sum_{post \in day} Exposure'(symbol, post, day)$$

5 Analysis of AG, GME using Our Methodology

We demonstrate our approach by applying the methodology on AG and GME stocks. AG is the ticker symbol of First Majestic Silver, a Canadian silver-mining company. At a late point during the GME incident, the GME stock reached a relatively high price, prompting people to start selling the stock in the thought that it reached its peak. This resulted in the GME stock losing its value. Some retail investors were looking to pivot as GameStop started losing momentum, and began targeting the silver stock AG, hashtagged *#silversqueeze*. The stock's price increased drastically from about 14 to 22 dollars, roughly 60%. However, unlike the GME event, this trend did not last for long and after about 5 days the AG stock price returned to its normal value.

5.1 Data

We used the subreddit WallstreetBets [7], and utilized PushShift API [5] in order to extract all the posts published from 25/01/2021 until 30/01/2021. Each comment object contains 26 attributes, we chose the attributes most relevant to our implementation:

- Post id - unique id of the submission
- SelfText - the textual content of the comment
- Title - author's title, only exists for the root comment (the initial post)
- Score - the number of upvotes minus downvotes for the comment
- URL - permalink, refers to the discussion itself
- Created_UTC - time of comment submission

We also added two more calculated fields for each comment:

- Tree depth of the comment
- Sentiment score of the comment, ranges between -1 and 1. We utilized VADER (Section 5.2.1).

5.2 NLP Techniques

5.2.1 Sentiment Analysis

We utilized VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [8]. We used the compound score provided by VADER to evaluate the sentiment of the comment. The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric when a single uni-dimensional measure of sentiment for a given sentence is needed.

We display the compound scores in our output graphs to manually examine if there is a correlation between the sentiment of the comment and the underlying action described in it (Buy/Hold/Sell, see Section 5.2.2). Red symbolizes a positive sentiment, blue symbolizes a negative sentiment, and gray symbolizes a neutral sentiment.

For example, in Figure 2 the root is grey, meaning a neutral sentiment, and contains the action *Hold* which makes sense since holding is a neutral action. On the other hand, a different blue node in the same tree represents the comment: *"16 year old here with only $1000 earned from shovelling shit. Can only think of one wise thing to do with all hard earned cash. GME to the moon"*. This contradicts the theory of a connection between the sentiment and the action since the comment motivates to buy the stock - a positive action, while the content is tagged as negative, probably due to the negative language. Another example is a red node representing the comment: *"We HOLD AND WE ALL GET FILTHY RICH! Avoid temptation to sell! Selling only allows them to cover their shorts cheaper, and slows the squeeze down"*, which clearly targets the action *Hold* but is tagged as having a positive sentiment.

After going over about 40 posts we came to the conclusion that there is no consistent connection, however, a more extensive research into this is needed in order to reach a definitive conclusion.

5.2.2 Named Entity Recognition (NER)

We used spaCy en_core_web_lg model [6], where *en* stands for the language the model was trained on (English), *core* refers to the type of capabilities the model has: vocabulary, syntax, entities, and vectors. *Web* refers to the genre of the data the model was trained on: written text, including blogs, news, and comments. Finally, *lg* stands for *Large*, referring to the model's size: 741 MB. The model was used to recognize entities in the comment's text. First, we filtered the entities according to the Organization entity label (ORG) and then we compared them to a pool of all publicly traded stock symbols [4]. The stock tickers are from the NYSE, NASDAQ, and AMEX.

5.3 Implementation Notes

We wanted to compute the exposure of AG and GME during a timeframe of a specific week. We constructed a tree graph from every post published during that time, utilizing networkX [3] (see examples in Figures 2, 3, 4). The node size is determined by the score of the comment compared to the scores of other comments in the same post. For each day we constructed a tree graph given all the posts published during that day. Using BFS traversal we computed the local popularity vector of the given discussion and the exposure of the stock symbol for the given day. In order to validate if a symbol is part of the comment we utilized an NER technique as explained in Section 5.2.2. We output the graph for each post whose exposure $E \geq 5$.

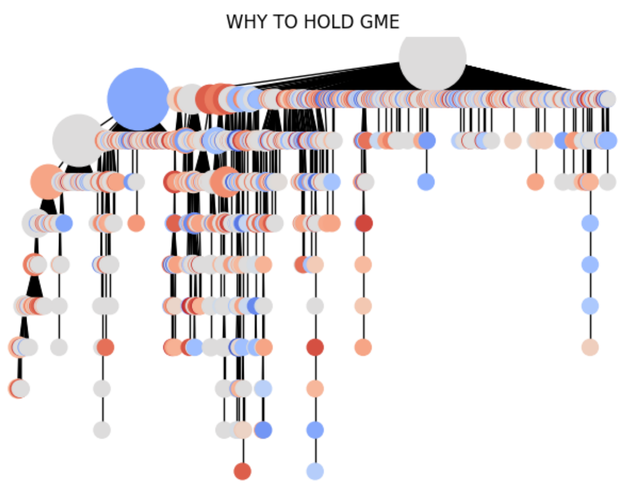


Figure 2: Exposure for GME in 25/01/2021 $E = 73$

Figure 3: Exposure for GME in 25/01/2021 $E = 91$

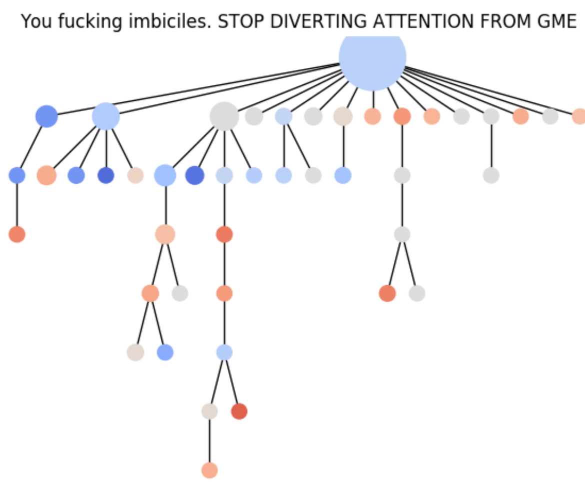


Figure 4: Exposure for GME in 27/01/2021 $E = 9$

5.4 Comparative Statistics

Tables 1, 2 describe statistics of the WSB posts extraced during the timeframe we analyzed for the AG and GME stocks. *#Related* stands for the number of posts related to GME ($E > 0$) and *%Related* stands for the percentage of posts related to GME out of the total number of posts. Max $Exposure'$ is the maximum $Exposure'$ over all posts during the indicated day.

Date	*Exposure*	Max *Exposure'*	#Related	%Related	#Posts
25/01/2021	7000.26	183.0	3627	25.65%	14139
26/01/2021	4621.70	259.0	1895	22.61%	8381
27/01/2021	16740.03	505.0	11138	25.89%	43022
28/01/2021	39005.22	763.0	23870	14.56%	163972
29/01/2021	26272.92	623.0	12774	11.18	114260

Table 1: WSB statistics on GME stock per day during the timeframe that we analyzed.

Date	*Exposure*	Max *Exposure'*	#Related	%Related	#Posts
25/01/2021	3.0	1.0	3	0.021%	14139
26/01/2021	2.0	1.0	2	0.024%	8381
27/01/2021	57.47	48.0	10	0.023%	43022
28/01/2021	133.14	8.0	126	0.077%	163972
29/01/2021	71.27	8.9	74	0.045%	114260

Table 2: WSB statistics on AG stock per day during the timeframe that we analyzed.

According to these statistics we can see an increase in the number of related posts both for AG and for GME, indicating an overall increase in overall exposure and therefore suggests the existence of a social trend. A comparison between Google's search trends during the same timeframe for AG and GME supports our findings (Figure 5). The peak in AG's interest occurred on February 1st, while our exposure statistics show that the post with highest exposure was on January 27, a few days earlier. A possible explanation could be that the increase in exposure for AG in WSB prompted the interest in the stock, causing higher popularity in Google's searches. In GME's case on the other hand, the peak in interest occurred on the 28th, in exact accordance with our exposure measures.

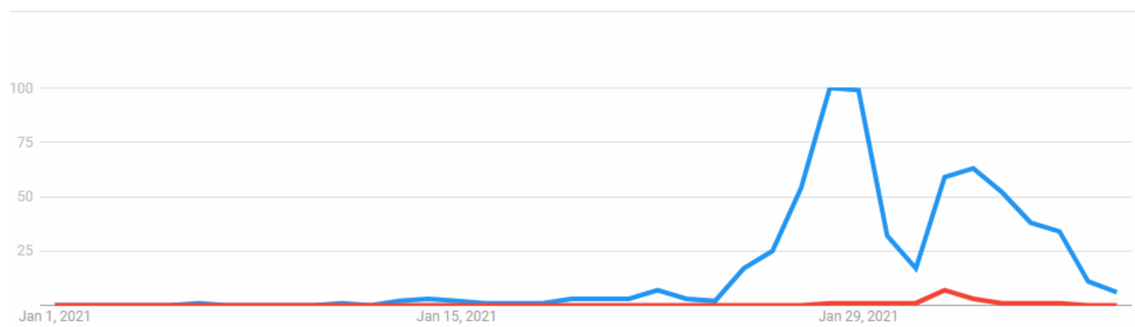


Figure 5: Google's interest over time graph for GME (blue) compared to AG (red). Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

We also measured the total exposure for the AG and GME stocks during the analyzed period (Figures 6, 7, 8, 9).

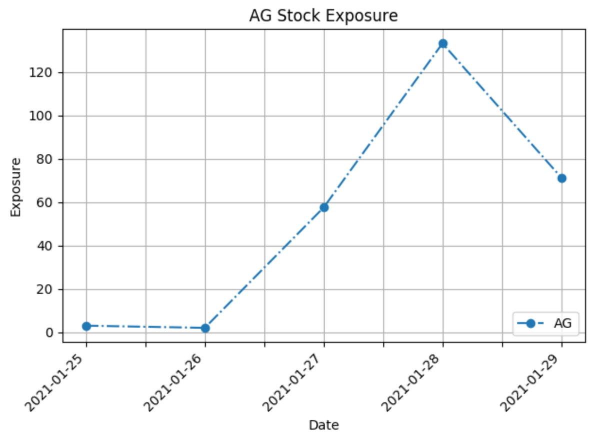


Figure 6: AG stock exposure

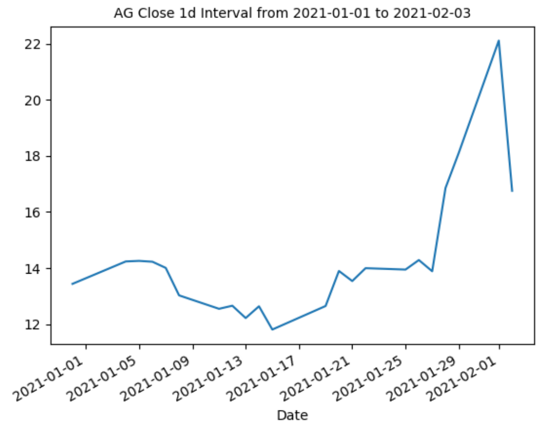


Figure 7: AG stock close price

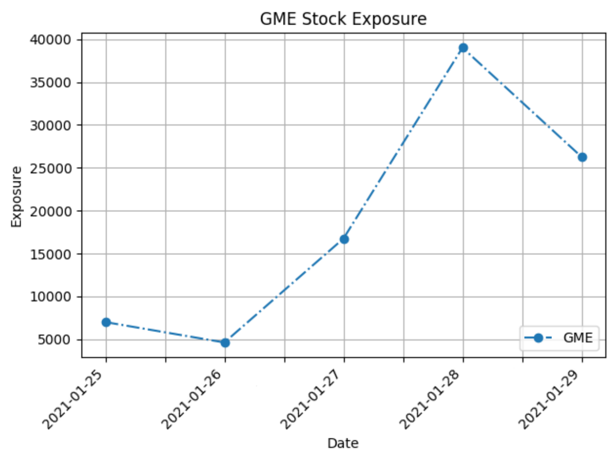


Figure 8: GME stock exposure

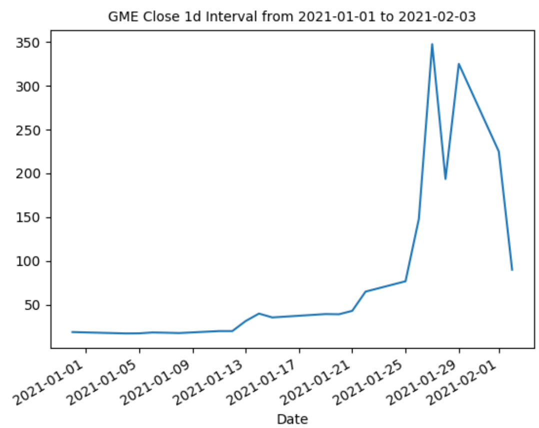


Figure 9: GME stock close price

By measuring the exposure over the predefined time interval we reach the following Spearman correlation measures between the exposure and the stock price:

- AG: 0.99
- GME: 0.50

In this particular analysis we can see that our approach points to a trend in both cases, however the scale of exposure could determine the duration of the trend (persistence over time) and whether or not a social cascade is created.

6 Conclusion

We believe that our approach summarises the information contained in the text differently from baseline approaches such as counting the total number of mentions, thus, giving a good estimation about social trends. Furthermore, by examining the collected data from our analysis, our approach predicts a trend before Google Trends. Since we were limited in computational power, we believe that our work can benefit from a larger-scale study that will apply our approach to more stocks and timeframes. Moreover, we believe that further research should be done with regards to the correlation between the sentiment of a comment and the action described in it. Perhaps using a different sentiment analysis method will help avoid tagging comments containing bad language as having a negative sentiment, since as we've shown in several examples- bad language does not always indicate a negative intention of the speaker. We believe that a feature-based deep learning model can help achieve greater sentiment classification accuracy by capturing nuances from context.

References

[1] Gamestop. https://www.gamestop.com/. Accessed: 2021-07-30.

[2] The gamestop story: how a group of investors on reddit gave wall street a wild week. https://theprint.in/theprint-essential/the-gamestop-story-how-a-group-of-investors-on-reddit-gave-wall-street-a-wild-week/595181/. Accessed: 2021-07-30.

[3] networkx python package. https://networkx.org/. Accessed: 2021-07-30.

[4] publicly traded stocks. https://github.com/shilewenuw/get_all_tickers. Accessed: 2021-07-30.

[5] Pushshift api. https://github.com/pushshift/api. Accessed: 2021-07-30.

[6] spacy en_core_web_lg documentation. https://spacy.io/models/en#en_core_web_lg. Accessed: 2021-07-30.

[7] Subreddit wallstreebbets. https://www.reddit.com/r/wallstreetbets/. Accessed: 2021-07-30.

[8] Valence aware dictionary and sentiment reasoner. https://github.com/cjhutto/vaderSentiment. Accessed: 2021-07-30.

[9] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[10] D. Bradley, J. Hanousek Jr, R. Jame, and Z. Xiao. Place your bets? the market consequences of investment advice on reddit's wallstreetbets. *The Market Consequences of Investment Advice on Reddit's Wallstreetbets (March 15, 2021)*, 2021.

[11] H. K. Sul, A. R. Dennis, and L. Yuan. Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3):454–488, 2017.

[12] C. Wang and B. A. Huberman. Long trend dynamics in social media. *EPJ Data Science*, 1(1):1–8, 2012.