Tracking Trending Stocks via Reddit Wallstreetbets

Section 1 Group 6
Khristion Lambert, Shijun Wei, Zhe Wang

Executive Summary

Since the start of 2021, the performance of some stocks has been abnormal, *Gamestop (GME)* is one of them. The reason for the pump and dump of these stocks may be the first time caused by individual investors.

A lot of investors, not only individual investors but also some institutional investors are eager to know the possible cause of the abnormal performance of the stocks. With that goal in mind, our team built a pipeline to collect data from the Internet and stored the data into the SQL database. The data collected was from the famous forum Reddit by using API. Besides, the web-scraping technique was also implemented to obtain the performance of the stock during the same trading period from Yahoo Finance. All the data was stored in the SQL database for further analysis.

By changing the filter in the pipeline, we were able to collect all the comments about certain stocks and their performance into one SQL table. With the implementation of the pipeline, investors can keep a close eye on the activities related to their positions and even guide their actions accordingly.

Introduction

“To the moon!” If you did not catch the reference, this is a common phrase shouted from the hilltops during the unprecedented rise of stocks early in 2021. During this time，we found many people jumping into the investment realm of stocks and options. If those people got in at the right time, they made tremendous profits, but if they were too late they

might have had significant losses. The reason so many investors were able to gain profit was due to the short squeeze on stocks. With so many people buying stocks it essentially caused the price of stock to skyrocket. Many of the people who trade stocks are legitimate business owners and they utilize strategies from various sources, one of those being the Wallstreetbets subreddit on Reddit.

Reddit is a social news aggregation/discussion forum where registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "communities" or "subreddits", which cover a variety of topics such as news, politics, science, movies, etc. Within these subreddit communities there has been one that has been experiencing some very unusual behavior that has made a lot of people very wealthy in a short amount of time. This subreddit is known as Wallstreetbets, which has been around since January 2012. In this channel the focus is on stock and option trading. Most recently, in January 2021, members of this channel engaged in a strategy known as a short squeeze of American video game retailer GME and other securities. Approximately 140% of GME's public shares had been short sold, and the rush to buy shares to cover those positions as the price rose caused it to rise ever farther. At the height, on January 28, the short squeeze caused GameStop's stock price to reach a pre-market value of over $500 per share, nearly 30 times the $17.25 share price at the beginning of the month. During this time, the price of other heavily shorted securities also increased. Naturally, this information went viral on various social media platforms, most notably Twitter. This sent the public in a frenzy where people from everyday traders to people with no prior

experience jumped in to make serious profits. As analysts we wanted to take a closer look into this market to understand what was going on during this time. Particularly we wanted to see if there was any kind of correlation between the number of comments talking about these stocks, the trade volume, and value of the stock.

We found that there were some general trends that led to this rise and as analysts we wanted to uncover these patterns. By digging deeper, we hoped to find a correlation between the number of people talking about these stocks and the trade volume and in effect the value of the stocks. We anticipated that these trends would hold constant, but we found that this was simply an anomaly that was unpredictable. Although this was a rarity, we thought it still might be interesting to understand the behavior of investors during this time.

Web-Scraping

First, we extracted stock data for GameStop through Beautifulsoup web scraping technique from Yahoo Finance. We chose this website because it provides comprehensive financial news, data and commentary including stock quotes, press releases, financial reports, and original content. This website also updates financial data on a daily basis; therefore, we can always obtain the most recent data by rerunning our code. In the project we requested the most recent 100 rows of data and stored the information in a table for analysis.

	Date	Open	High	Low	Close	Adjusted Close	Volume
0	2021-03-12	275.00	295.50	262.27	264.50	264.50	25760700.0
1	2021-03-11	241.64	281.50	232.60	260.00	260.00	28186000.0
2	2021-03-10	269.43	348.50	172.00	265.00	265.00	71361900.0
3	2021-03-09	217.71	249.85	208.51	246.90	246.90	38725800.0
4	2021-03-08	154.89	210.87	146.10	194.50	194.50	63424800.0
...	...	...	...	...	...	...	...
95	2020-10-23	15.05	15.38	14.55	15.00	15.00	6507300.0
96	2020-10-22	14.20	15.87	14.19	14.91	14.91	16212200.0
97	2020-10-21	13.90	14.42	13.80	14.10	14.10	5361900.0
98	2020-10-20	14.03	14.14	13.67	13.86	13.86	6604000.0
99	2020-10-19	13.44	14.50	13.38	13.91	13.91	13169100.0

100 rows × 7 columns

Figure 1. The stock prices collected from Yahoo Finance

Extracted stock data includes Dates, open price, highest price, lowest price, close price, adjusted close price and trading volume for GME daily to prepare for more analysis in the future. By exploring the HTML script in Yahoo Finance, we found each day of stock price is embraced by a class tag called ‘BdT Bdc($seperatorColor) Ta(end) Fz(s) Whs(nw)’, so we iterated through information contained in all of those tags and extracted data that we need to store into this table.

API

There are two API that can collect data from Reddit: PRAW and Pushshift. We tried both and found only Pushshift can collect data in a certain time range that we chose. We used Pushshift for the data collection. And the time range of the collected data is from Jan.1st.2021 to Mar.14th.2021. Using Pushshift, we can change the search criteria:

Before, which a timestamp of the beginning date is needed

After, which a timestamp of the ending date is needed

query, which we can type in the keyword(s) to look for in submissions

sub, which we can type in the Subreddit we want to search in.

Using the search criteria, data can be collected automatically for all the comments of each stock in a certain time range we are interested in.

We successfully collected all the comments which mentioned the popular stocks like GME, AMC in the subreddit: Wallstreetbets. There are 143949 records about GME during the time period we collected. And there are 61 variables that were collected for each comment. All the data is restored in JSON format as shown in Figure 2.


```
{
    "data": [
        {
            "all_awardings": [],
            "allow_live_comments": false,
            "author": "JayStax17",
            "author_flair_css_class": null,
            "author_flair_richtext": [],
            "author_flair_text": null,
            "author_flair_type": "text",
            "author_fullname": "t2_zrnom",
            "author_patreon_flair": false,
            "author_premium": false,
            "awarders": [],
            "can_mod_post": false,
            "contest_mode": false,
            "created_utc": 1615296419,
            "domain": "i.redd.it",
            "full_link": "https://www.reddit.com/r/wallstreetbets/comments/m16cll/yolod_90_of_my_portfolio_into_gme_at_130_not_a/",
            "gildings": {},
            "id": "m16cll",
            "is_crosspostable": true,
            "is_meta": false,
            "is_original_content": false,
            "is_reddit_media_domain": true,
            "is_robot_indexable": true,
            "is_self": false,
            "is_video": false,
            "link_flair_background_color": "#349e48",
            "link_flair_css_class": "profit",
            "link_flair_richtext": [
                {
                    "e": "text",
                    "t": "Gain"
                }
```


Figure 2. The data collected from Reddit API in a JSON format

We transformed the JSON format to Pandas dataframe as shown in Figure 3. Among all the variables, we chose 9 key variables:*"Post ID", "Title", "Url", "Author", "Score"," Publish Date", "Total No. of Comments", "Permalink", "Flair"* for further analysis.All the extracted data is saved in .csv files in local drives.

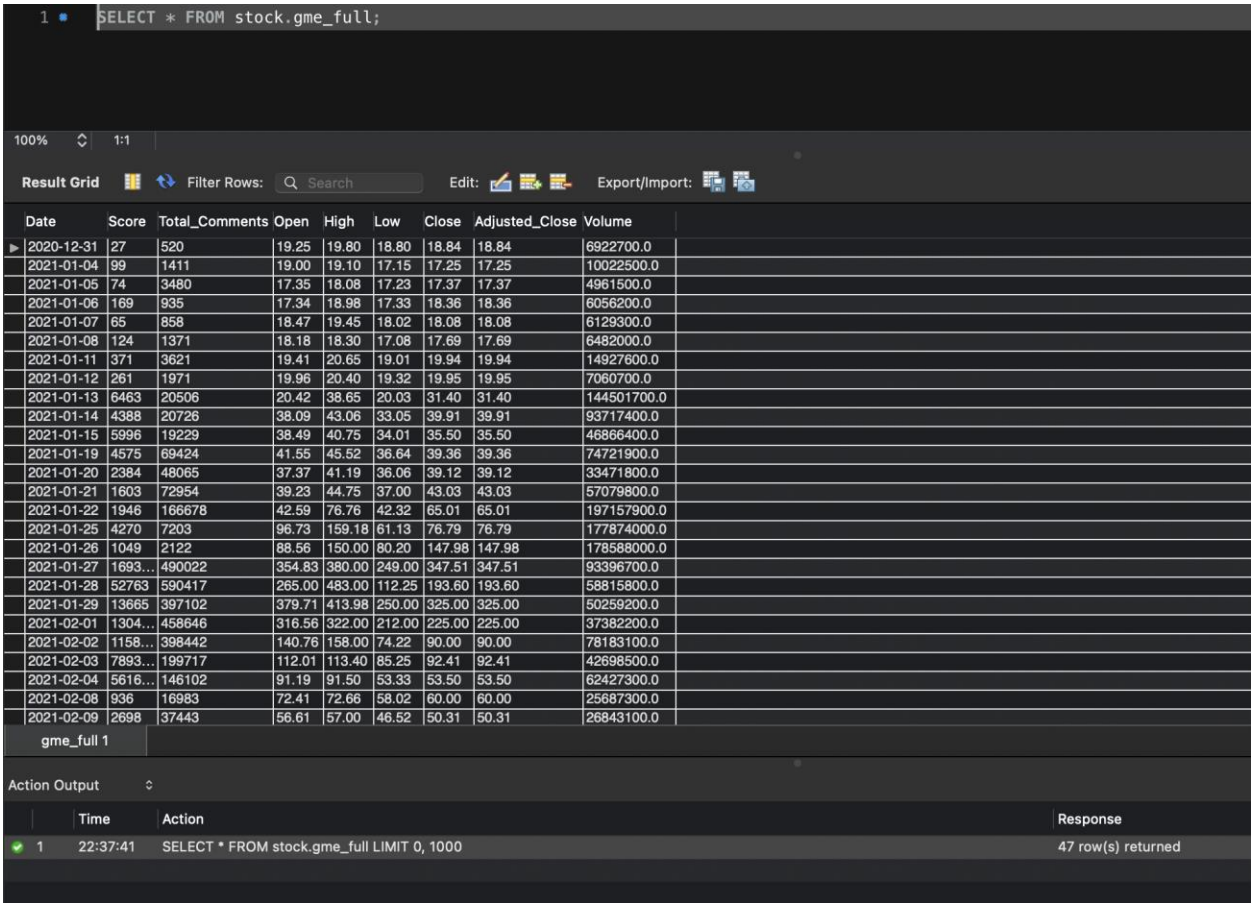
Post ID	Title	Url	Author	Score	Publish Date	Total No. of Comments	Permalink
ko145e	GME to 420.69, but only if we make it happen. ...	https://www.reddit.com/r/wallstreetbets/commen...	stevenconrad	1	2020-12-31 16:05:29	5	/r/wallstreetbets/comments/ko145e/gme_to_42069...
ko1bnp	What would make GME shorts win?	https://www.reddit.com/r/wallstreetbets/commen...	dluther93	1	2020-12-31 16:18:03	0	/r/wallstreetbets/comments/ko1bnp/what_would_m...
ko1kck	Not sure how reliable a random comment in Cohe...	https://pbs.twimg.com/media/EqmGLzSXEAArjcz?fo...	MilitaryBeetle	1	2020-12-31 16:32:49	36	/r/wallstreetbets/comments/ko1kck/not_sure_how...
ko1ttx	🚀🚀🚀🚀How have we been so fucking blind? GME is ...	https://www.reddit.com/r/wallstreetbets/commen...	WSBProfitProphet	1	2020-12-31 16:49:32	0	/r/wallstreetbets/comments/ko1ttx/how_have_we_...
ko1xxb	GME is the Rockets 🚀🚀🚀🚀	https://www.reddit.com/r/wallstreetbets/commen...	WSBProfitProphet	1	2020-12-31 16:56:35	11	/r/wallstreetbets/comments/ko1xxb/gme_is_the_r...

Figure.3 The data collected from Reddit in a panda dataframe.

SQL

We then wrote a snippet in python to connect with MySQL in the local server and created a database ‘stock’. We created a table in the database named “gme_full”. Having the table ready, we need to define the tables with all the column names as well as the primary key(s). “Date” is defined as the primary key for the table. As Shown in the Figure. 4, there

are 9 fields in the table which have all the information we collected from the Internet.



```
1 • SELECT * FROM stock.gme_full;
```

100% 1:1

Result Grid Filter Rows: Search Edit: Export/Import:

Date	Score	Total_Comments	Open	High	Low	Close	Adjusted_Close	Volume
2020-12-31	27	520	19.25	19.80	18.80	18.84	18.84	6922700.0
2021-01-04	99	1411	19.00	19.10	17.15	17.25	17.25	10022500.0
2021-01-05	74	3480	17.35	18.08	17.23	17.37	17.37	4961500.0
2021-01-06	169	935	17.34	18.98	17.33	18.36	18.36	6056200.0
2021-01-07	65	858	18.47	19.45	18.02	18.08	18.08	6129300.0
2021-01-08	124	1371	18.18	18.30	17.08	17.69	17.69	6482000.0
2021-01-11	371	3621	19.41	20.65	19.01	19.94	19.94	14927600.0
2021-01-12	261	1971	19.96	20.40	19.32	19.95	19.95	7060700.0
2021-01-13	6463	20506	20.42	38.65	20.03	31.40	31.40	144501700.0
2021-01-14	4388	20726	38.09	43.06	33.05	39.91	39.91	93717400.0
2021-01-15	5996	19229	38.49	40.75	34.01	35.50	35.50	46866400.0
2021-01-19	4575	69424	41.55	45.52	36.64	39.36	39.36	74721900.0
2021-01-20	2384	48065	37.37	41.19	36.06	39.12	39.12	33471800.0
2021-01-21	1603	72954	39.23	44.75	37.00	43.03	43.03	57079800.0
2021-01-22	1946	166678	42.59	76.76	42.32	65.01	65.01	197157900.0
2021-01-25	4270	7203	96.73	159.18	61.13	76.79	76.79	177874000.0
2021-01-26	1049	2122	88.56	150.00	80.20	147.98	147.98	178588000.0
2021-01-27	1693...	490022	354.83	380.00	249.00	347.51	347.51	93396700.0
2021-01-28	52763	590417	265.00	483.00	112.25	193.60	193.60	58815800.0
2021-01-29	13665	397102	379.71	413.98	250.00	325.00	325.00	50259200.0
2021-02-01	1304...	458646	316.56	322.00	212.00	225.00	225.00	37382200.0
2021-02-02	1158...	398442	140.76	158.00	74.22	90.00	90.00	78183100.0
2021-02-03	7893...	199717	112.01	113.40	85.25	92.41	92.41	42698500.0
2021-02-04	5616...	146102	91.19	91.50	53.33	53.50	53.50	62427300.0
2021-02-08	936	16983	72.41	72.66	58.02	60.00	60.00	25687300.0
2021-02-09	2698	37443	56.61	57.00	46.52	50.31	50.31	26843100.0

gme_full 1

Action Output

	Time	Action	Response
1	22:37:41	SELECT * FROM stock.gme_full LIMIT 0, 1000	47 row(s) returned

Figure. 4 The Saved Table ‘gme_full’ in SQL database ‘stock’.

Conclusion

In general, we want to hit our target to provide business insights to investors who are looking for a data-driven method to analyze and measure what correlates with the performances of their stocks. In this report, we acquired financial information specifically for GameStop stock from Yahoo Finance, and then concatenated them with GameStop comments data retrieved from Wallstreetbets subreddit on Reddit. In the end, we inserted

all the cleaned data into a SQL table for easy retrieval to be further analyzed. Those data we collected can be used to explore the correlation between comments from Wallstreetbets and stock's performance. For example, whether the number of comments is related to the fluctuation of stock's trading volume? Moreover, by doing a sentiment analysis on the comments content, we might be able to understand whether commentor's attitude towards GME in a series of days will impact the price or trading volume of the stock itself.