Capstone Project – A look into the Reddit/WSB data

Author: Hui Cao
Date: 4 April 2021

INTRODUCTION

Recently, a few stocks become the top picks of the Reddit/WSB platform, an online finance discussion forum where people exchange opinions and ideas on stock markets. Within a short period of time, share price of those stock rockets to historical high before pluming into the normal low level and soared once more, maintaining an extreme volatility. Because of the incredibly huge gain when bets are correct leveraging on options, Reddit/WSB has become a must-visit place before executing transactions related to those stocks.

In this report, based the data from Reddit/WSB, the author tries to discover patterns, stories that might be interesting to investigate. For instance, questions can be: what are the most popular topics or stocks? When do people normally involved in the discussions? Does threads volume change with time? Any correlation of threads volume and share prices concerned?

DATA

Dataset is a csv file downloaded from Kaggle. There are 8 columns and 44,192 rows in the dataset. After wrangling, the author kept 5 columns that are most relevant to the analysis. Namely 'title', 'score', 'id', 'url', 'comms_num', 'created', 'body', 'timestamp'.

	title	score	id	url	comms_num	created	body	timestamp
0	It's not about the money, it's about sending a...	55	l6ulcx	https://v.redd.it/6j75regs72e61	6	1.611863e+09	NaN	2021-01-28 21:37:41
1	Math Professor Scott Steiner says the numbers ...	110	l6uibd	https://v.redd.it/ah50lyny62e61	23	1.611862e+09	NaN	2021-01-28 21:32:10
2	Exit the system	0	l6uhhn	https://www.reddit.com/r/wallstreetbets/commen...	47	1.611862e+09	The CEO of NASDAQ pushed to halt trading "to g...	2021-01-28 21:30:35
3	NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...	29	l6ugk6	https://sec.report/Document/0001193125-21-019848/	74	1.611862e+09	NaN	2021-01-28 21:28:57
4	Not to distract from GME, just thought our AMC...	71	l6ufgy	https://i.redd.it/4h2sukb662e61.jpg	156	1.611862e+09	NaN	2021-01-28 21:26:56

METHOTHOLOGY

The author first wrangled the data by dropping less relevant columns, then added three columns of 'date', 'weekday' and 'hour' to explore the pattern of the posts (or 'threads').

Next, the author moved on to see the volume of the post change by counting the number of posts grouped by date, weekday, and hour, to see when people were more active in the platform and how the posts volume changed over time.

Word Cloud is useful to have a good glance at popular topics, the author performed word cloud based on title and content of all posts respectively. In addition, world clouds of top 10 most scored posts were also generated to explore the differences, if there was any.

The author wondered if the volume of the post was correlated to the price of the top pick stocks, namely AMC and GME, therefore, by introducing yfinance package, share price of those stocks were generated and plotted. However, the share price range of the two stocks are different, to better reflect the movement, the percentage change was used in the plot.

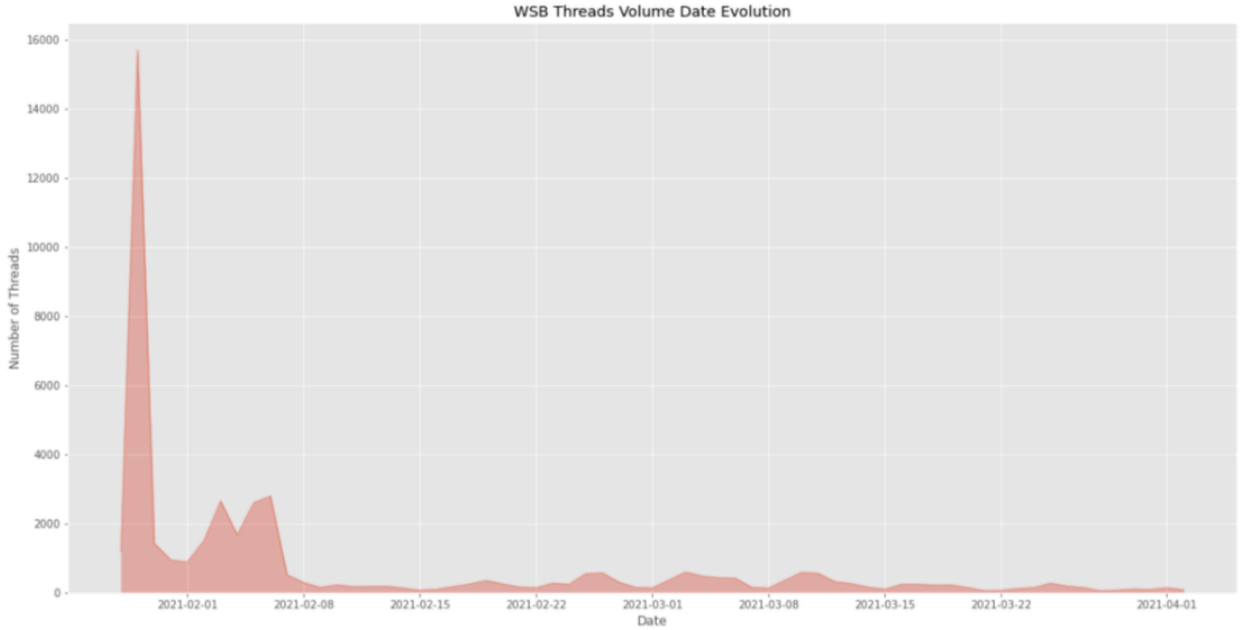
Next, to analyze the relation between the stocks and the volumes of posts, the author first plotted the two variables to visualize the relation, an extreme outlier of post volume was found, this outlier was excluded to better capture the relation. Then correlation and P-value were calculated to quantify the linear relation of stock movement and volume of posts. Furthermore, correlation of the two stocks was also checked to see similarity of those two stocks.

Last, Kernelling was also used to test whether the relation of stock price and posts volume is polynomial, in order to predict pattern underlying those two variables using Machine Learning techniques.

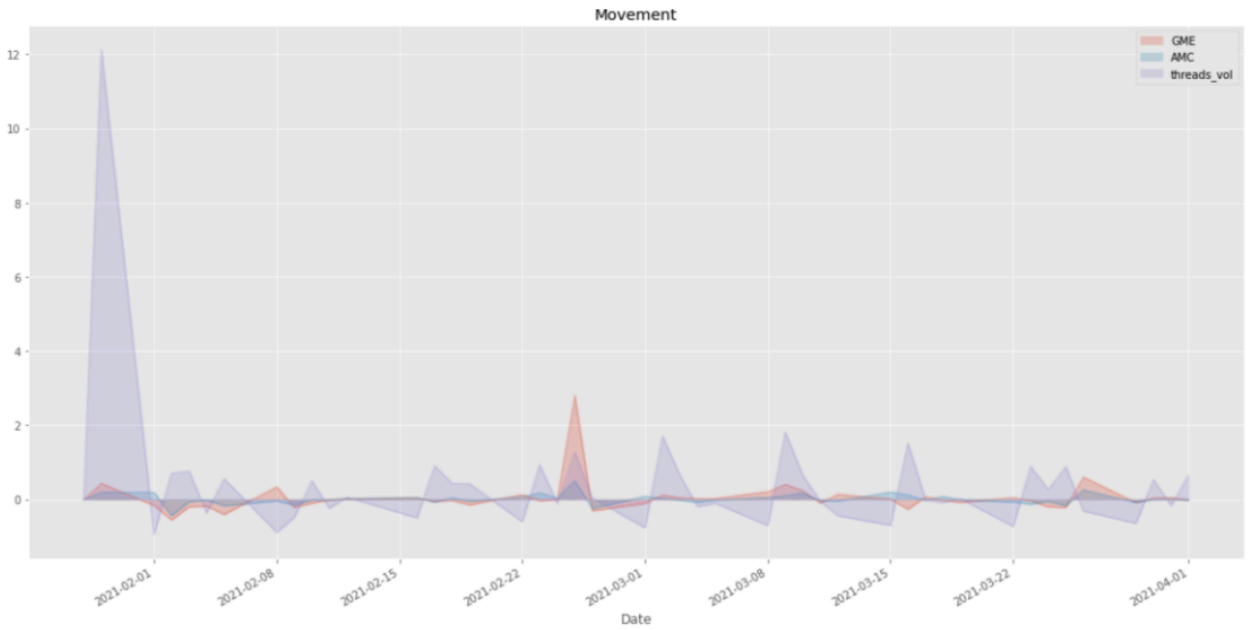
RESULTS

Based on the line of reasoning discussed above, the author has come across the below findings.

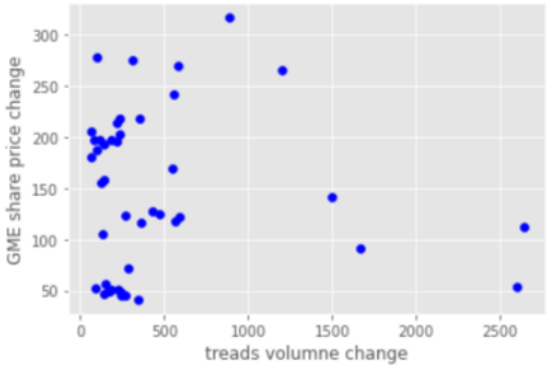
1. The number of posts soared on 29th January 2021, and plumed quickly before rising gain in later months, however, never retained the pick again.

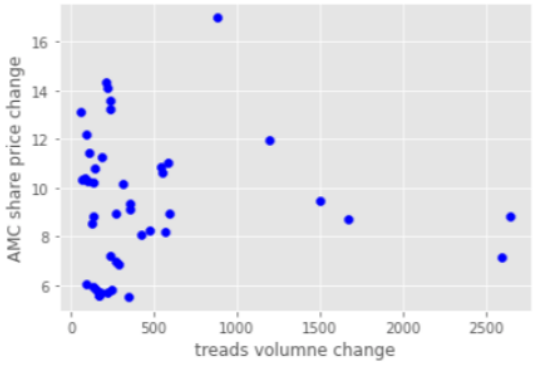


Nevertheless, the surge of posts number did not always share the rise of price for the top-picks, namely ‘GME’ and ‘AMC’. As demonstrated in the below chart recoding percentage movement of share prices and post volume change.



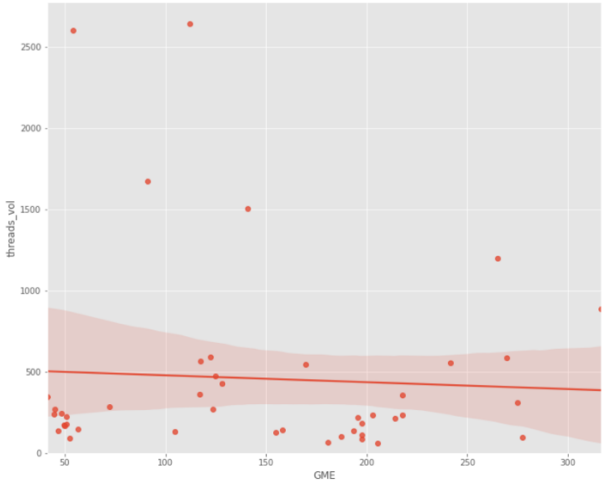
2. In addition, price change of AMC and GME did share a lot of similarities, with regard to posts/threads volume changes. The Pearson coefficient and P-value of GME share price to post volume was -0.06 and 0.7 respectively (-0.005 and 1 for AMC), suggesting a week negative linear correlation.



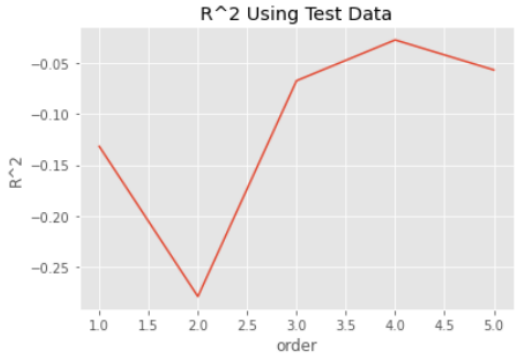


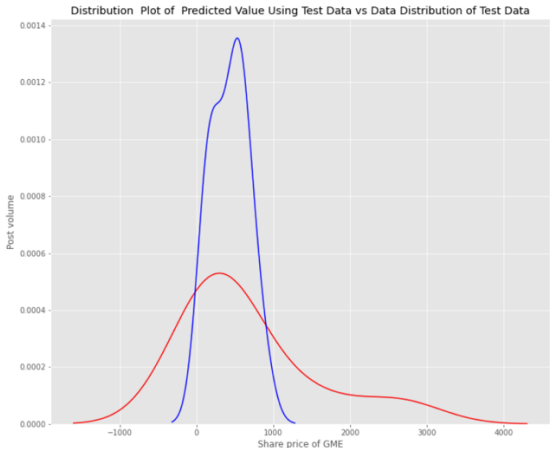

```
The Pearson Correlation Coefficient is -0.005283399090993009  with a P-value of P = 0.9728473346644443
```

	GME	AMC	threads_vol
GME	1.000000	0.879604	-0.056245
AMC	0.879604	1.000000	-0.005283
threads_vol	-0.056245	-0.005283	1.000000

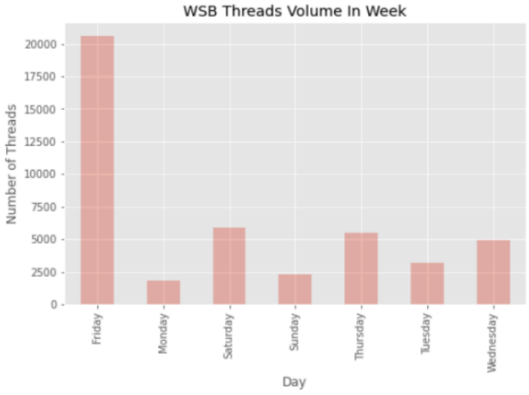


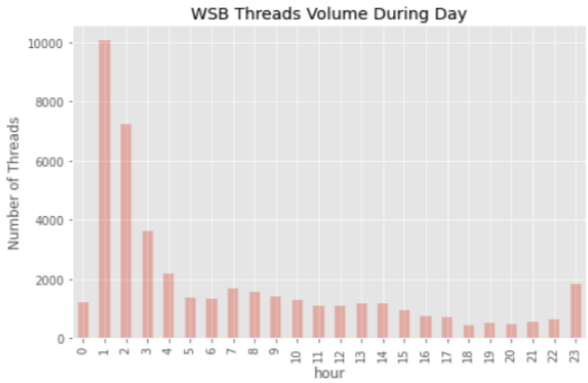
3. Next, it could be that the relation of price and post volume is polynomial. A test of order was conducted, resulting a 4th order regression. However, applying a 4th order regression did not seem to a good predictor – the distribution plot of predicted value using test data and real test data did not match very well.





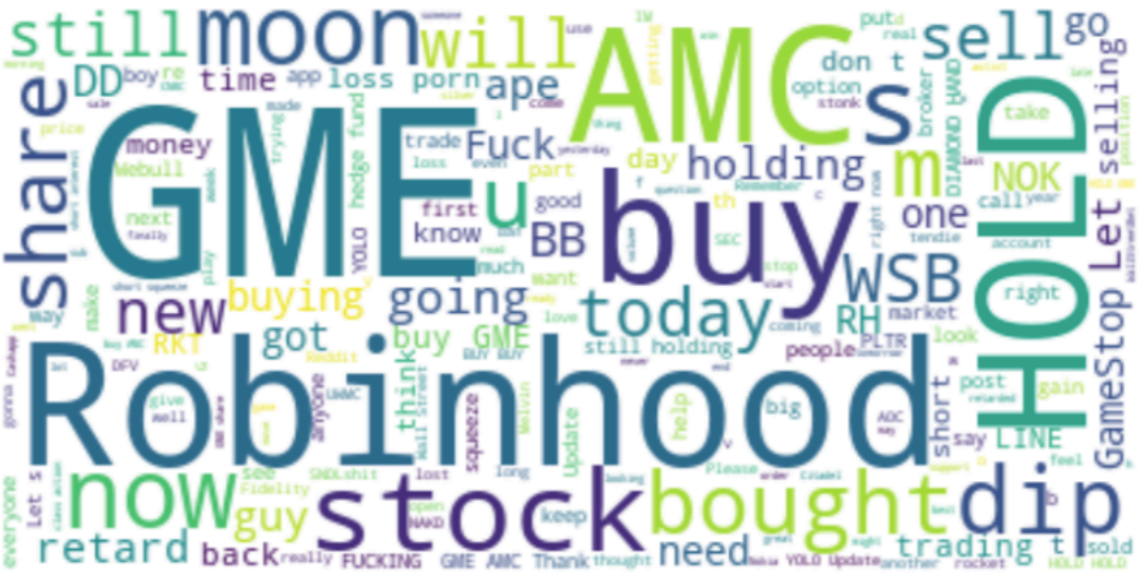
4. It looks like a lot of people were participating in the party of "yolo", as shown above in the posts volume. A good look at the time of the posts created could be a good indicator of behavioral observation of people active on Reddit/WSB. As shown below, most people were active on Friday, which make sense, considering that people are mostly having more free time on Friday and that Friday is the last day of the week when market is open. Besides, most people post threads at 1-2 a.m. in the morning, perhaps that most people have time in the early Friday morning discussing the strategy to execute the next day.





5. Last, what exactly did people discuss? Below four graphs were generated based on the title and content of the posts/threads, and that of the top 10 posts/threads ranked by score of the posts. As shown below, the words were quite similar among those 4 word clouds. The most popular words were, to name a few, ' GME', 'AMC', 'Robinhood', 'Stock', 'buy' , 'moon', 'BB', 'share', 'hold', 'YOLO', 'hedge', 'fund'.

   Word Cloud by Title of the posts



Word Cloud by Body of the posts

Word Cloud by Title of the top10 posts ranked by score

DISCUSSION & CONCLUSION

As demonstrated above, the author note that the posts number surged at peak in late January of 2021 then was kept at a moderate level, suggesting a short-term mania of the heated ‘treasure-hunting’ behavior. The most popular topics in the Reddit/WSB

were mostly around the top picks, which were ‘GME’ and ‘AMC’, against the hedge fund, etc. To the author’s great surprise, the correlation of the stock price and the volume of the posts were quite weak, at least no apparent linear relation. A higher order regression did seem work either. Therefore, the author concluded that there was no significant link between the share price of ‘GME’ or ‘AMC’ with the enthusiasm of the people in the forum, therefore it would be impossible to predict price of those stocks based on the number of the posts.

Lastly, it seems that people normally surfed the forum Friday which had the peak volume of posts created, and people were mostly active very late at night (or very early in the morning), meaning that most people use free time after work to discuss investment strategies with regards to those stocks.