UNIVERSITY OF MILAN-BICOCCA

REDDIT SUBMISSIONS AND TWEETS ANALYSIS DURING GameStop SAGA

LUCA BALLARATI	867488
FRANCESCO GREGORI	889206
FRANCESCO OLIVIERO	812292

Summary

REDDIT SUBMISSIONS AND TWEETS ANALYSIS DURING’GAMESTOP SAGA'

INTRODUCTION

The stock meme craze started years ago, in a little corner of the internet called r/WallStreetBets, a Reddit community with a penchant for low-level humor and a detective understanding of the stock market. This page was launched in 2012 by Jaime Rogozinski, at the time a technology consultant for the Inter-American Development Bank.
It brings together people who share investment ideas and thoughts in the same community. It has become particularly popular, reaching almost 3 million users, starting from March 2020, a period coinciding with the lockdown situations caused by the Covid-19 pandemic. This growth has been accompanied by the increased interest of small savers in the world of equity investments and the increase in the popularity of ’free’ trading platforms, such as Robinhood.

The reason it is well known, however, is another: it played a fundamental role in GameStop's short squeeze, causing large losses in a few days for some US companies and short sellers. In the stock market, a short squeeze is a rapid rise in the price of a stock due primarily to excess short selling rather than underlying fundamentals. A short squeeze occurs when there is a lack of supply and excess demand for the stock due to short sellers having to buy stocks to cover their short positions.

The American video game retailer GameStop, like all companies that do not invest in digital, does not have a great future. After the outbreak of the pandemic, the situation for society worsens. We start talking about bankruptcy, so much so that large investment companies such as Melvin Capital begin to bet on the company's misfortunes by selling shares short. Short selling, or short selling, is a financial transaction that consists in the sale of securities not directly owned by the seller, but borrowed against payment, with the intent of obtaining a profit following a movement bearish on a stock exchange .

At the beginning of December 2020, the GameStop company recorded an operating loss of 63 million dollars in the third quarter. Shares on the New York Stock Exchange plummeted nearly 20% with a value of $ 13.66 per share.
The following month the company appoints three new directors: Ryan Cohen, founder of the Chewy e-commerce platform, and two of his collaborators. The latter, during 2020, had already shown interest in the company by purchasing shares.

The announcement elicits some early chatter on the Reddit r/WallStreetBets page. Discussions are born as memes, that is, as jokes. In such discussions it is proposed to buy GameStop shares in bulk, which are very cheap. The goal is to hold these stocks while waiting for the value to go up (*diamond hand*).

On the other hand, the large investment companies continue with the downward bet, continuing to shorting in order to counteract the increase in the value of the shares.

The meme becomes reality, a lot of people start buying shares in the reference company, removing them from the market, causing the price to rise accordingly.
While the value of the shares rises disproportionately, the shorting companies find themselves forced to close everything as soon as possible, buying back the short position, leading to a consequent rise in the stock.
Over 71 million debts (short positions) were generated by these loans, despite GameStop does not have so many shares on the market.

At the end of January Elon Musk tweets “GameStonk!!!”, a clear link to the reddit r/WallStreetBets forum. In the same day, Chamath Palihapitiya, a high-profile venture capitalist, via twitter, says he is investing in GameStop.
The GME stock rose 140% to $ 354.83 per share.
Major short sellers, such as Citron Research and Melvin Capital close their positions with a significant loss.
Robinhood and other platforms restrict transactions for GME.

All this captures the attention of legislators and political figures. The SEC even comes into play.
This great phenomenon represented an extraordinary event since, for the first time, the speculative bubble does not come from large operators, but from many individual and coordinated actions of small users.

The purpose of this project is to analyze whether the social networks Reddit and Twitter have really influenced this incredible increase in the value of the shares of the GameStop company, now on the verge of bankruptcy.
The interest is to look for some relationship between the different submissions made on the r/WallStreetBets page by users and the fluctuation of the GME share price, and some other link between the various tweets, of reference persons or not, and the same increase in the shares of the famous American video game retailer.
Through the construction of a graph database, we want to reconstruct the social network of users (reddit on the one hand, twitter on the other) who actively participated, through comments and personal opinions posted on the reference social networks, in the short squeeze of GameStop. The aim is therefore also to identify the social relationships between the various users, highlighting the relevant ones, thus finding those people (retail investors) who had the greatest influence in the great event described.
We also want to determine which group of people orchestrated the entire operation.

DATA ACQUISITION

The data necessary to carry out the analysis described above relates to the submissions and corresponding comments made on the r/WallStreetBets subreddit, to the tweets and to the

various historical quotations on the New York Stock Exchange (NYSE) of the GME share of GameStop.
squeeze occurred is considered: indicatively for the entire month of January 2021.

Submissions and tweets are posts made by users on the Reddit and Twitter ’social networks’ respectively.
Reddit is an American social news aggregation, web content rating and discussion website. Registered members submit content to the site such as links, text posts, images and videos, which are then rated up or down by other members. Posts are organized by topic in user-created message boards called "community" or " subreddit ", covering topics such as news, politics, religion, science, movies, video games, music, books, sports, fitness, cooking, pets and sharing. of images.
Twitter is an American microblogging and social networking service where users post and interact with messages known as "tweets".

For the historical data concerning the GME stock, it was decided to take advantage of the large yahoo finance database.

In order to extract submissions, tweets and quotes, according to the search parameters, the same methodology is used: the data is downloaded via API.
An Application Programming Interface (API) is a standardized and secure interface that allows applications to communicate and work with each other. It allows you to request data from a third party provider so that you can use this information as and when needed.

REpresentational State Transfer (REST) API defines how applications can communicate over HTTP to transfer information efficiently and quickly.

For some services the use of the related API is free, for others limited access is allowed and for others it is necessary to pay a fee.
For the data acquisition the Python language was used, mainly supported by the Pandas library executed in several Jupyter notebooks.
The main container for the data is the Pandas dataframe, which is completely similar to a table.
The dataframe represents the entire collection of documents, corresponding to each row of the table. Instead, the fields make up the columns.

Twitter data acquisition

Twitter provides a series of APIs to be able to query its services. Such bees require different levels of authorization. Accounts with an ’elevated’ level can use more services.
Academic Researcher Account: Gives qualified academic researchers high access and advanced functionality, including access to the full archive search endpoint, a monthly Tweet limit, and improved filtering capabilities with filtered stream and search endpoints recent.

The aim of the project is to get the tweets sent in January 2021, so you need an academic researcher account to be able to retrieve them from the Twitter API. The student account does not have these privileges, so it was necessary to go another route.
It was decided to use the *snscrape library,* which allows you to retrieve tweets data without having a specific account and without limitations.
The operation is simple: once the ’twitter’ module of the library has been set up, it is sufficient to invoke the appropriate function that starts the ’scraping’ of the data by passing the keywords that must be contained in the tweet text and the time interval within which do the research.
Given the volume of data to be recovered, the download lasted several hours and was carried out at different times and with different search parameters (different keywords and different time intervals).
At the end of the download, the data, retrieved in JSON format from the library, are inserted into a data frame Pandas and stored in a csv file.
This file was saved in order to have the data available in offline mode, for subsequent data management and manipulation operations, in particular for the data cleaning and data quality phases, using the Python Pandas library.

At the end of the activities, the data was saved on the database.
The reference file for the above operations is ’GameStopTwitter.ipynb’

The reference attributes of a tweet are listed below:

- **date** = date and time the tweet was made
- **tweetId** = tweet id
- **userId** = user id who tweeted
- **display_name** = name user on display
- **text** = text of the tweet (may contain external links, emoticons...)
- **reply_count** = count of replies to the tweet
- **like_count** = like count of the tweet
- **retweet_count** = retweet count
- **quote_count** = count of quote tweets or tweets. Lets say something besides the retweet
- **retweted_tweet**
- **quotedTweet**

Reddit data acquisition

Similarly to what was done for Twitter, there are special libraries for Reddit as well.
After a careful search on the web to find the best solution, it was decided to use the Pushshift library which allows you to easily retrieve submissions and comments.
Submissions are the messages that a user posts, while comments are messages intended to comment on a submission or another comment.
To further simplify the process of data recovery from Reddit, Python code was used which allowed to further mask the operations performed by the Pushshift library.

Submissions and comments were downloaded in JSON format and saved on different files, since it was not possible to carry out the operation in a single tranche. The files were then merged and saved in a single file (one for submissions and one for comments) in csv format.

Data verification and manipulation operations were carried out and finally they were saved on the database.
The reference file for the above operations is ’GameStopReddit.ipynb'

The reference attributes of a submission are listed below:

- **total_awards_received** = silver, gold or platinum award for a submission. represents an evaluation system. For example, the silver award is a symbol of appreciation, the gold award associates appreciation and reward credits for the submission author.
- **author** = user who posted the submission
- **author_premium** = membership
- **created_utc** = date and time the submission was made
- **flair** = submission label
- **is_video** = boolean value to indicate whether it is a video or not (0 = no, 1 = yes)
- **num_comments** = number of comments received
- **score** = number of upvotes - number of downvotes
- **self_text** = post body
- **title** = post text
- **is_submission** = boolean value to indicate whether it is a submission or not (0 = no, 1 = yes)
- **is_op** = original poster
- **subreddit_subscribers** = subscriber volume of the subreddit (r/WallStreetBets) at the time of submission
- **id** = id submission
- **upvote_ratio** = upvote rate
- **no_follow** = Reddit initially classifies each post as nofollow.
  This discourages webmasters or paid third parties from spamming their links on the site. This doesn't mean that every Reddit backlink is always nofollow.

The reference attributes of a comment are listed below:

- **total_awards_received**
- **author** = author of the comment
- **author_premium**
- **created_utc**
- **flair**
- **is_video**
- **num_comments**
- **score**
- **self_text**
- **title**

- **is_submission**
- **is_op**
- **subreddit subscribers**
- **upvote_ratio**
- **no_follow**
- **id** = comment id
- **link_id** = reference submission id to which the comment was made
- **parent_id** = id comment to which the sub comment was made

Data acquisition of GME share prices

Historical data regarding GME share prices on the NYSE stock exchange was retrieved using Yahoo Finance's yfinance library.
With a simple call to a specific function it is possible to retrieve the data of interest, that is, day, opening value, closing value, max, min and volume. It is also possible to group the data by shorter or longer intervals, for example, among others, by hour and by day.
The data is returned directly in a Pandas dataframe, therefore they can be easily stored in a csv file.
The reference file for the above operations is ’GameStopStockQuotes.ipynb’

The reference attributes of a GME quote are listed below:

- **open**
- **high**
- **low**
- **close**
- **volume**
- **dividends**
- **stock splits**

DATA STORAGE

For the subsequent data storage it was necessary to choose the type of DBMS most suitable for the purpose of the project.
The choice fell on Neo4J, therefore a graph database.

Neo4J

Neo4j is a graph database software open source developed entirely in Java .

are often faster than relational databases at associating datasets, and map the structures of object-oriented applications more directly. They scale more easily, to large amounts of data, and do not require the typical and costly *join operations.*

The choice of this type of database is justified by the fact that a graph database allows to model the relationships between nodes.
One of the objectives of the project is to find and visualize relationships between the users of the social networks that have most influenced the whole affair.
In the specific case, the nodes are represented by users and messages, while the relationships indicate the actions that users and messages do or undergo.
It was not possible to relate the two worlds of social networks involved, namely Twitter and Reddit ; therefore, the relationships exist only within the actors of the two social networks.

The modeling chosen for Twitter is as follows:

- (User) - [writes] -> (Tweet)
- (Tweet) - [mentions] -> (User)
- (User) - [mentioned] -> (User)

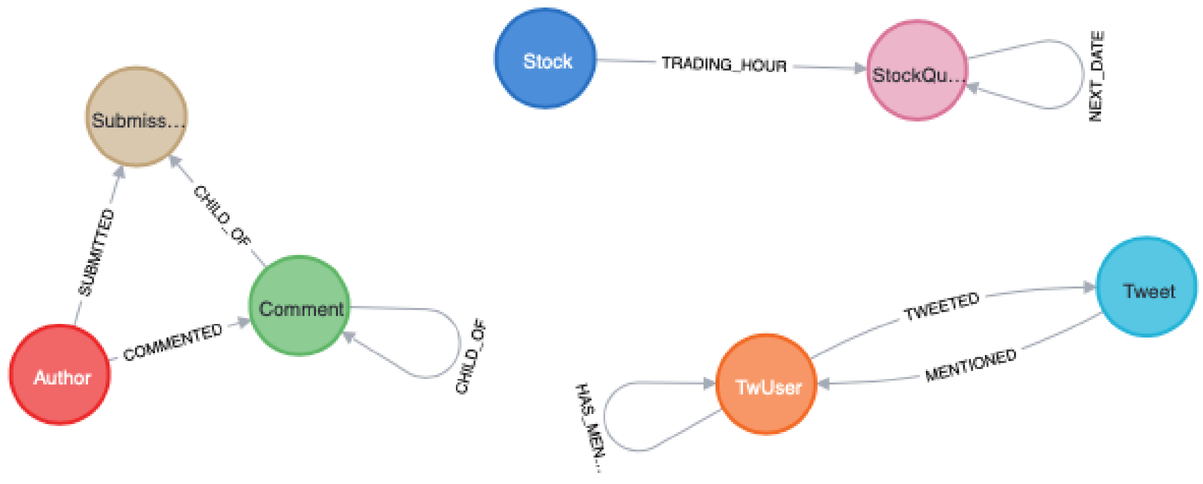
The modeling for Reddit is as follows:

- (Author) - [comment] -> (Comment)
- (Author) - [send] -> (Submission)
- (Comment) - [son] -> (Submission)
- (Comment) - [child] -> (Comment)

For the sake of completeness, the data relating to the GME share were also stored in the database.
The ease and speed of retrieval of the related information, as well as the small volume of data, mean that the necessary data can be taken directly from the network without going through the storage of the same on db.
Their modeling on db is as follows:

- (Stock) - [trading_ hour] - > (StockQuote)
- (StockQuote) - [next_ date] - > (StockQuote)

The resulting scheme on Neo4J is therefore the following

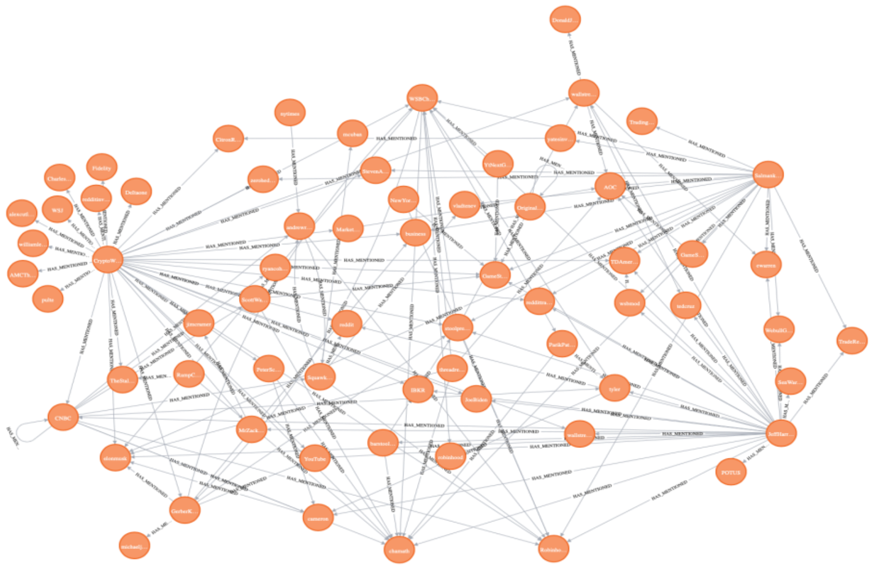


The entire datasets recovered during the data acquisition phase were loaded. At a later stage, only the data of interest will be selected for the analyzes.

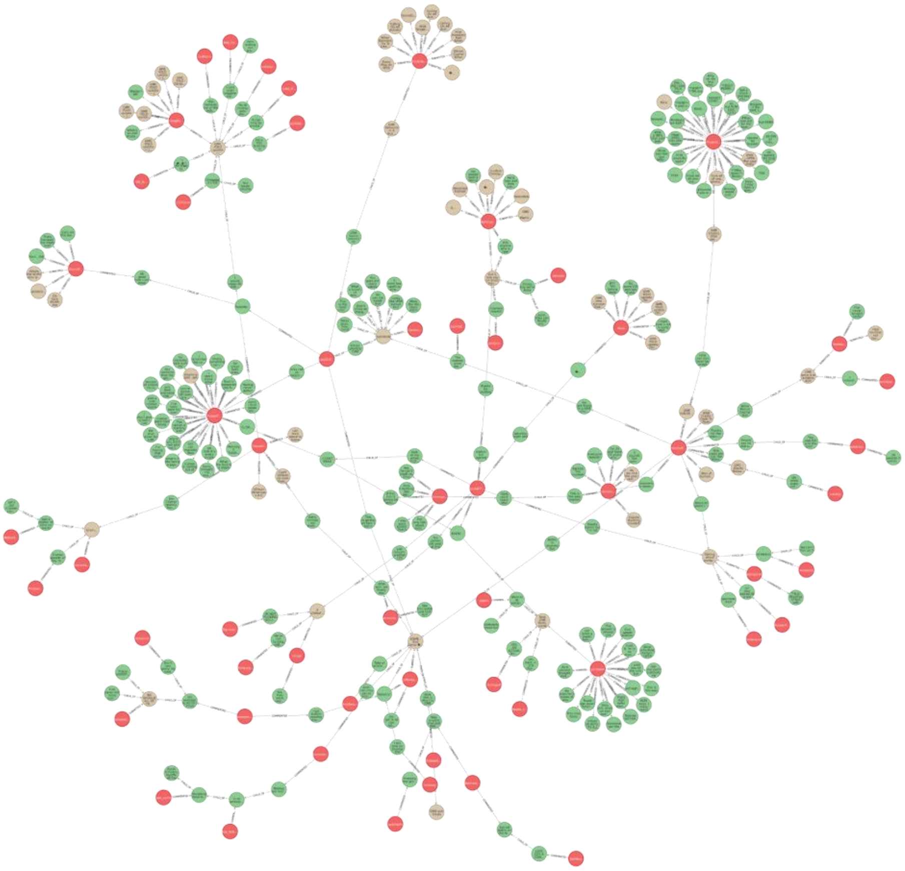
The subsequent queries were carried out in order to obtain only the messages containing specific keywords involved in the operation to increase the price of GME shares.

Once the data was loaded, it was also possible to perform specific queries, in Cypher language, to see if there were any connections between users that suggest a possible joint action to cause the increase in the value of the GME stock.
For this purpose, with regard to Twitter, users were searched for connected to each other through the mentions made in the tweets and having a high number of 'has_mentioned' relationships, both incoming and outgoing.
The calculation of this number was done thanks to the Neo4J degree centrality algorithms.

The query result is displayed below:



As for Reddit, links between users (Author as indicated on db) were searched through the relationships between the same users and the Comment and Submission nodes.
The resulting graph is the following, with the red Author nodes:



Installation of the db

The choice of where to create the database is not trivial.

The goal is to have a database on the cloud so that it can be reached by different users.

AuraDB version of Neo4j was initially used, but a problem immediately arose: the free version allows you to enter a maximum of 50k nodes and 175k relations. Too little for the volumes involved in the project.

The choice therefore fell on Neo4J sandbox, an environment always in the cloud with no limitations on the number of nodes and relationships, but with a limited duration, maximum 10 days.

The limit encountered with this version was that of having a too stringent timeout on transactions. This limit is manifested in the use of the libraries of procedures for the calculation of the degree centrality, for the Twitter nodes.

db was installed locally in a Docker container. With this solution it was possible to use the enterprise version of Neo4J and have no limitations, if not the physical limit of the machine.

Twitter data was uploaded to the local db, the degree centrality was calculated and then entered on the db in the cloud, in order to have the data available and updated for all users.

DATA PREPARATION

Before inserting the data into the database, it is necessary to perform some data preparation operations in order to have as much clean, complete and quality data as possible

Data integration

It is not possible to carry out a real data integration between the Twitter and Reddit datasets, since there are no fields that can be used as a link between the two social networks.

The data acquisition on Twitter was performed in several steps, producing three different csv files
A set union was made between the files obtained (concatenation in Pandas)
d at the moment that the data frames contain disjoint data, that is, one refers to the time interval 17-23 January, one refers to the interval 24-30 January and the last includes the overall period, but has been requested with other keywords.

No data integration even with regards to Reddit; dataframes containing submissions and comments have been left separate.
However, the columns containing the message text have been standardized:

- Submission: text contained in the ’title’ column, while the ’self_text’ column contains only a blank space
- Comment: inverse of the previous one, text in the ’self_text’ field and ’title’ containing blank space

‘title’ column was then used for both dataframes.

Data quality

Data quality checks and consequent changes were made on the dataframes, obtained during the data acquisition

Completeness

null values or empty string on two types of fields

- required fields: ’tweetId’ and ’text’ for Twitter, ’id’ and ’title’ for submissions and comments on Reddit
- unnecessary fields: all the rest

Actions: remove lines with null values for ’tweetId’ and ’text’ ('id’ and ’title’ for Reddit), insert empty string for the remaining fields.

Uniqueness

Twitter:

- Check for duplicate rows: a row is duplicated if it has identical values in the ’tweetId’ field

Reddit:

- Check for duplicate rows on Submissions and Comments on the ’id’ field: there are no duplicate rows

Actions: remove duplicate rows

Consistency

Twitter:

- Verification of the correctness of the ’date’ field, containing the date / time of the tweet, through the Pandas to_datetime function.
    - The field was found to be correct, as the function did not generate any errors; therefore, no further action was required

Reddit:

- Field ’created_utc’: contains the date / time in long
    - conversion into datetime: it did not generate an error, therefore it is believed that the field has a consistent value

Dataset tweet_df:

Total original dataset records: 1098653
Total records with missing text and tweetId: 105
Total dataset records after removal: 1098548
Total records with duplicate rows: 111 102
Total dataset records after removing duplicates: 1027319

Dataset submissions_df:

Total original dataset records: 18453
Total records with missing title and id: 0
Total dataset records after removal: 18453
Total records with duplicate rows: 0
Total dataset records after removing duplicates: 18453

Dataset comments_df:

Total original dataset records: 17787
Total records with missing title and id: 0
Total dataset records after removal: 17787
Total records with duplicate rows: 0
Total dataset records after removing duplicates: 17787

DATASET	N° RECORD ORIGINALI	N° RECORD CON ‘text’ E ‘tweetId’ NULL	N° RECORD DOPO PULIZIA	COMPLETEZZA
tweet_df	1098653	105	1027319	99,99%

DATASET	N° RECORD ORIGINALI	N° RECORD CON RIGHE DUPLICATE	N° RECORD DOPO PULIZIA	RIDONDANZA
tweet_df	1098548	71229	1027319	6,48%

DATASET	N° RECORD ORIGINALI	N° RECORD CON ‘title E ‘id NULL	N° RECORD DOPO PULIZIA	COMPLETEZZA
submissions_df	18453	0	18453	100%
comments_df	17787	0	17787	100%

DATASET	N° RECORD ORIGINALI	N° RECORD CON RIGHE DUPLICATE	N° RECORD DOPO PULIZIA	RIDONDANZA
submissions_df	18453	0	18453	0%
comments_df	17787	0	17787	0%

The new line and tab characters are then replaced with a space:

- for Twitter: 211466 lines involved
- for Reddit: involved 0 lines for Submissions and 3085 lines for Comments

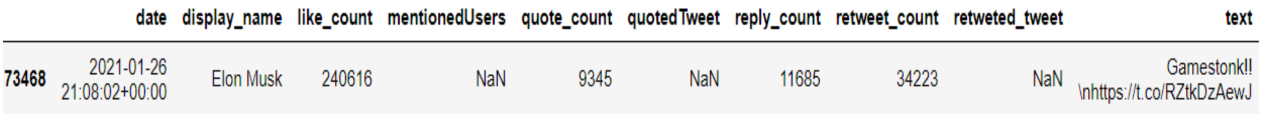
Finally, for all remaining lines, the null values of the fields are replaced with an empty string:

- Twitter: 2816520 fields involved
- Reddit: 512 fields for Submissions and 3554 for Comments

The modified dataframe is saved on CSV file to avoid losing the changes

A rather significant check was done: checking for the presence of a particular tweet that seems to have started the race to the top, referring to wallstreetbets, a subreddit of the social platform Reddit’

tweet_df[tweet_df ['username'] == ‘elonmusk’]



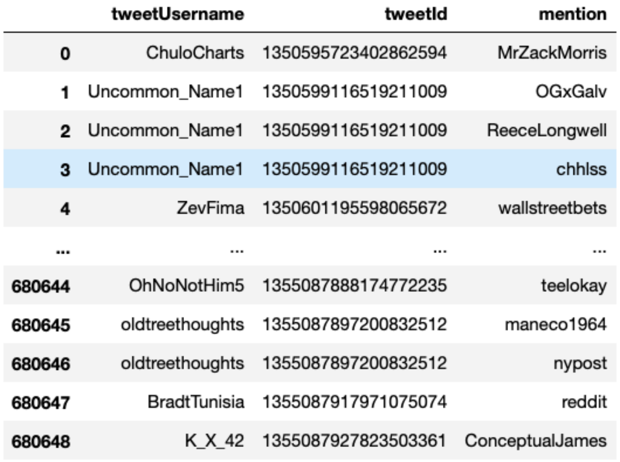
```
                   date display_name like_count mentionedUsers quote_count quotedTweet reply_count retweet_count retweted_tweet                    text
73468        2021-01-26   Elon Musk     240616            NaN        9345         NaN       11685         34223            NaN             Gamestonk!!
         21:08:02+00:00                                                                                                         \nhttps://t.co/RZtkDzAewJ
```

Data manipulation

Data manipulation operations were required for both Twitter and Reddit to extract information on the users involved

For Twitter, the users mentioned by the tweet text were extracted.
A mention to another Twitter user is identified by means of the @ character followed by a series of characters representing the user's name and terminated by a space. Therefore, with a special regular expression it was possible to extract all the users mentioned in a specific tweet.

It was possible to obtain a dataframe containing, for each row, the user who wrote the tweet, the identification of the tweet and the user mentioned.

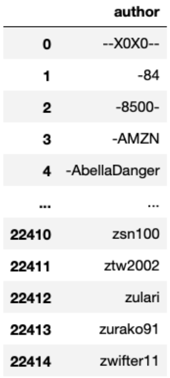


	tweetUsername	tweetId	mention
0	ChuloCharts	1350595723402862594	MrZackMorris
1	Uncommon_Name1	1350599116519211009	OGxGalv
2	Uncommon_Name1	1350599116519211009	ReeceLongwell
3	Uncommon_Name1	1350599116519211009	chhlss
4	ZevFima	1350601195598065672	wallstreetbets
...	...	...	...
680644	OhNoNotHim5	1355087888174772235	teelokay
680645	oldtreethoughts	1355087897200832512	maneco1964
680646	oldtreethoughts	1355087897200832512	nypost
680647	BradtTunisia	1355087917971075074	reddit
680648	K_X_42	1355087927823503361	ConceptualJames

Total records: 680649
Duplicate total records, i.e. also including the records to keep: 10601

Also as regards the mentions, duplicates are verified and removed, these are 5360 records, finally saved in CSV format
This results in 675289 users.

For the Reddit part, the ’Authors’ from submissions and comments are extracted, merged into a single dataset and subjected to cleaning of the inevitable duplicate records, since an Author can carry out both Submissions and Comments.
The total of single Authors is 22415.



	author
0	--X0X0--
1	-84
2	-8500-
3	-AMZN
4	-AbellaDanger
...	...
22410	zsn100
22411	ztw2002
22412	zulari
22413	zurako91
22414	zwifter11

Data enrichment

Neo4J’s Graph Data Science Library was useful as a data enrichment operation.
Specifically, the degree centrality algorithm was used to find the most ’popular’ nodes within the two graphs, that of Reddit and that of Twitter.
Degree centrality measures the number of relationships entering, leaving (or both) of a node, depending on the orientation of a relationship projection.
Each node of interest, in this specific case the TwUser nodes for Twitter and Author for Reddit, has been ’enriched’ with a new property, called ’degree', containing the result of the execution of the aforementioned algorithm.
The queries used are found in the ’GameStopTwitter.ipynb’ and ’GameStopReddit.ipynb’ files.

Text Analysis

A simple text analysis is necessary to understand which are the most common words in the texts of the messages sent by users on both social networks.
WorldCloud library, also in python, was then used to obtain the count of the occurrences of all the single words of each tweet, submission and comment. Subsequently, the most suitable ones were chosen to identify a message involved in the ’gamestonk’ operation.

For Twitter, the identified words are:

- gme
- gamestonk
- squeeze
- robinhood
- wallstreetbets
- wsb

The word GameStop was not used, as it was a search term for tweets during the scraping operation.

For Reddit, the following were identified:

- gme
- GameStop
- gamestonk
- squeeze
- robinhood
- everyone
- moon

in addition to the emoji 🚀.
wsb and wallstreetbets were not used as the recovered messages already belong to the subreddit wallstreetbets.

Users of the r/WallStreetBets subreddit use emojis or combinations of emojis, in code, to represent certain financial actions:

- 🚀 = indicates which shares they hope to "send to the moon" ie to rapidly increase the price

- 💎🙌 =‘*diamond hand’*, means to keep your investments and avoid selling while waiting for the price to rise
- 🧻🙌 =‘*paper hand’*, ranks investors who sell their shares before others
- 🐂 = represents the bulls on an investment, those who think it will increase in value ('bull gang')
- 🧸 = denotes those who are bearish on an investment ('bear gang')
- 🍗 = stands for " tendies ". It means that you will cash in your " tendies ", that is, you will realize your profits from an investment by closing your position.

The keywords identified above were used in the queries Cypher to select only the messages of interest from the database to then be analyzed and displayed. The reference files containing the aforementioned queries are ’GameStopRedditQueries.ipynb’and’GameStopTwitterQueries.ipynb’

TIME SERIES

Through the available databases, you want to go and graphically view whether there is actually a link between the posts made on Reddit and Twitter and the squeeze of the GME share price. In fact, one of the research questions aims to find a positive correlation, in the analyzed time frame, between the exchange of memes and discussions carried out on the subreddit r/WallStreetBets regarding the company GameStop and the trend of the share price. GME, and a positive correlation between the number of tweets posted on the social platform relating to investments in the video game company and changes in the share price.

GameStop’s share price appears to have been generated by this community on Reddit of small retail investors. Almost for fun, the idea of buying shares in the company in bulk was born. The excess demand would have caused the price to go up, as well as the fact that many short sellers would have had to close their positions.

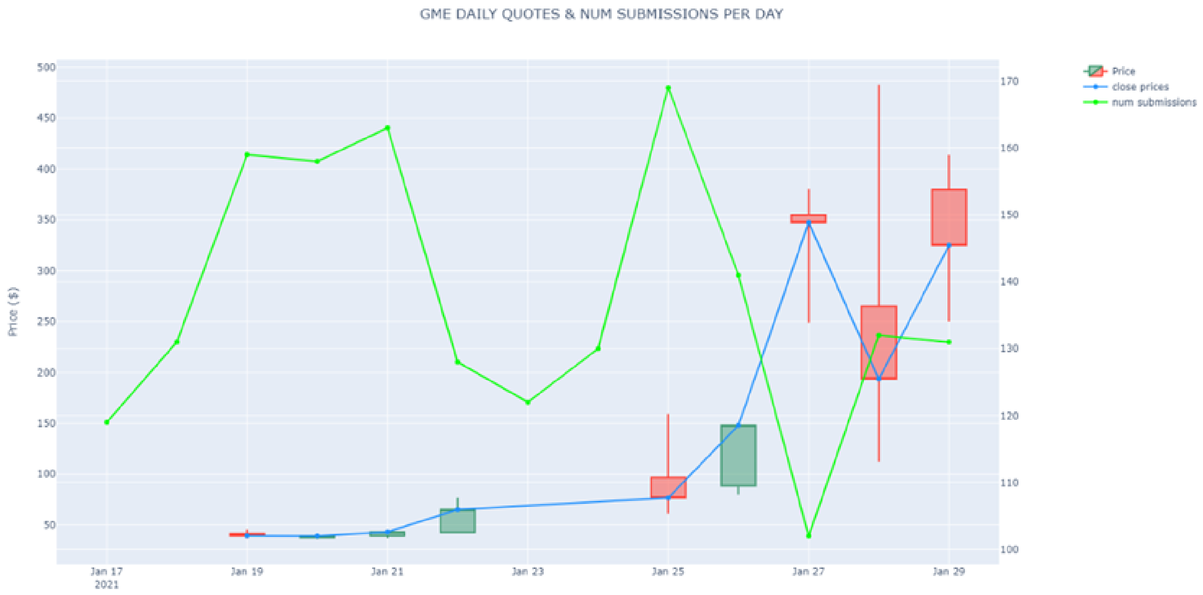
The social network of a platform like Reddit is very large. So many people have become aware of the possibility of being able to become rich and at the same time keep a company like GameStop alive, which probably represents something important for many users of the community.

After filtering submissions, submission comments and tweets by keywords, with reference to investments in GameStop, the number of these posts was calculated with daily frequency and hourly frequency, in the time interval between January 17, 2021 and 30 January of the same year. These data come from queries previously carried out using the Cypher language on the Neo4j software, and then downloaded in CSV format. References are made to ’reddit_submissions.csv', ’reddit_comments.csv', ’tweets.csv', ’twitter_users.csv’ and ’stock_quotes.csv'.

After a careful analysis it was decided to use the count of tweets per day and per hour, the count of submissions per day and finally the count of comments and submissions (sum) per day.

The frequency of submissions and comments per hour is too irregular to assess a possible link with the performance of the stock. This is due to the fact that the user of a subreddit is much lower than the user of Twitter, where they also comment on highly influential characters, which facilitate the visibility of a certain topic. There are therefore certain times of the day when no posts are registered on the platform.

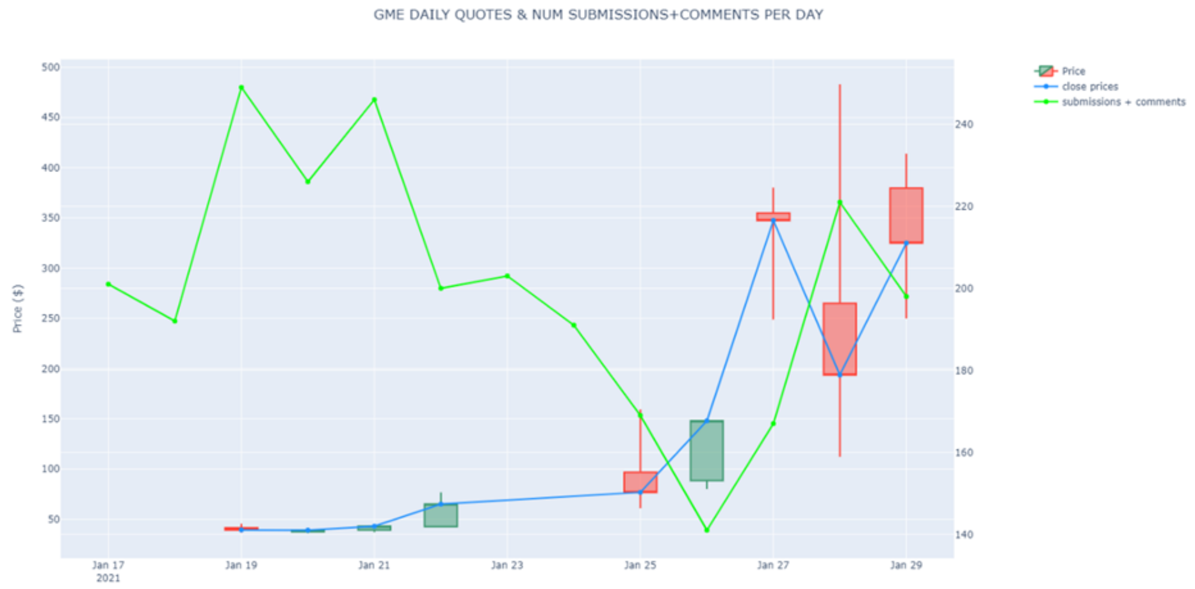
In the infographic below, through a candlestick plot, the price fluctuation of the single GME share is highlighted, together with the number of submissions per day.



As early as January 17, on Reddit, attention to the topic GameStop was alive. Towards the 19th of the same month, one of the highest peaks in the number of daily submissions is reached, regarding the situation of the video game retailer. It is assumed that the first big discussions about the idea of investment arose, but that still few users had acted by buying stocks. The price of the GME share is in any case increasing. There are still short sellers who are hoping for a downward trend and continue to borrow stocks.

The maximum post peak is identified 2 days before the maximum price reached by GME shares. This is quite plausible: the goal is to buy and then hold for the stock's value to rise. This is what happened, also thanks to the closing of the shorting positions ; in addition, the trend of the two trends is very similar.

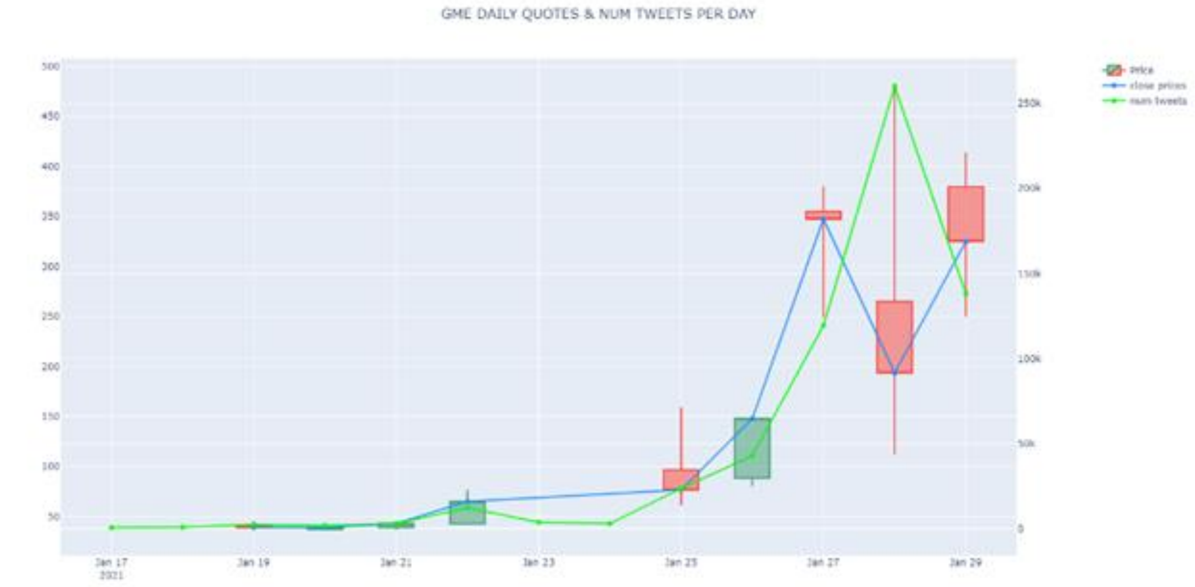
We also wanted to view the daily trend of GME prices together with the daily evolution of the number of submissions and comments taken together.



GME DAILY QUOTES & NUM SUBMISSIONS+COMMENTS PER DAY

The trend changes slightly: there is even more attention in the first days of the time horizon considered, after which there is a decline, and it ends with another peak following the great squeeze of the GME stock.

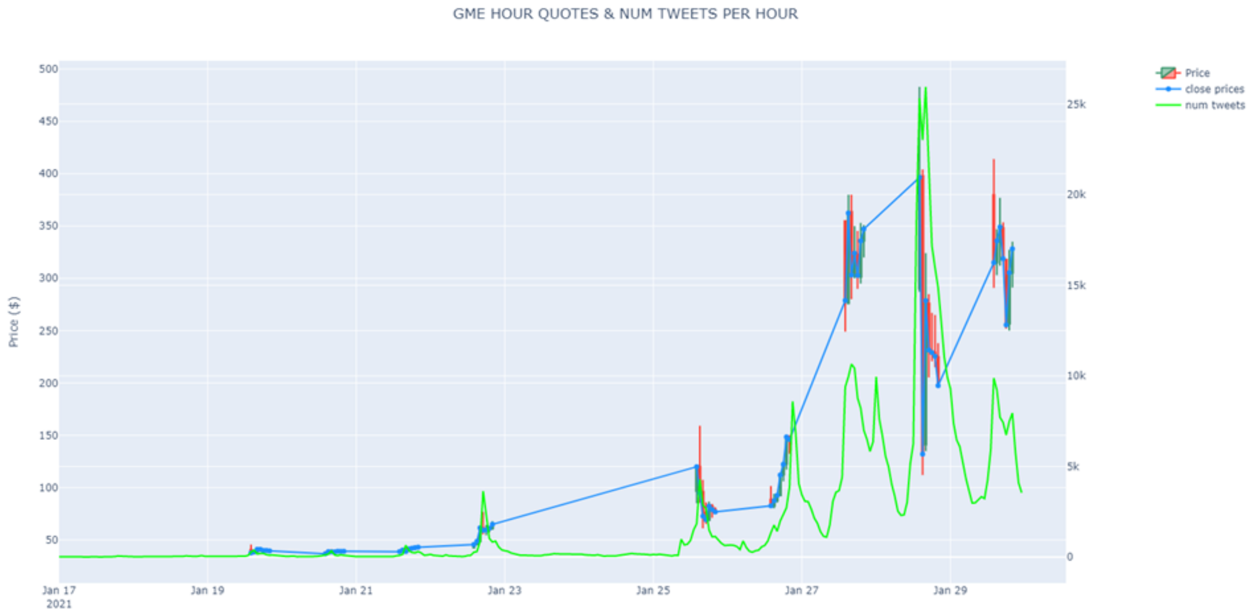
As for Twitter, the correlation with the share price trend seems even more evident.



GME DAILY QUOTES & NUM TWEETS PER DAY

As the number of tweets increases, so does the price of the title, or vice versa. There seems to be a really strong link between the two different variables. The post ’Gamestonk!!!’ of January 26 of Elon Musk, he certainly caught the attention of a large audience. It can be seen, as from that moment, the number of daily tweets has increased 4 times as much until it reaches its maximum peak, 2 days later. In addition, the popular Tesla CEO is no stranger to the power of being able to move the price of certain stocks, such as Tesla, or cryptocurrencies, such as bitcoin, up or down with a single tweet.

By viewing the data hourly, despite the probable numerous irregular fluctuations, the situation does not change. The relationship between the two variables is always present, it seems even more fortified.



However, it should be noted that Twitter and Reddit are completely different platforms. Twitter is populated by the most influential people in the world: those who have the most visibility are those who have the most followers. It is also seen as an information / gossip tool, not as a forum on certain topics. There are also pages and accounts of many media that share news.

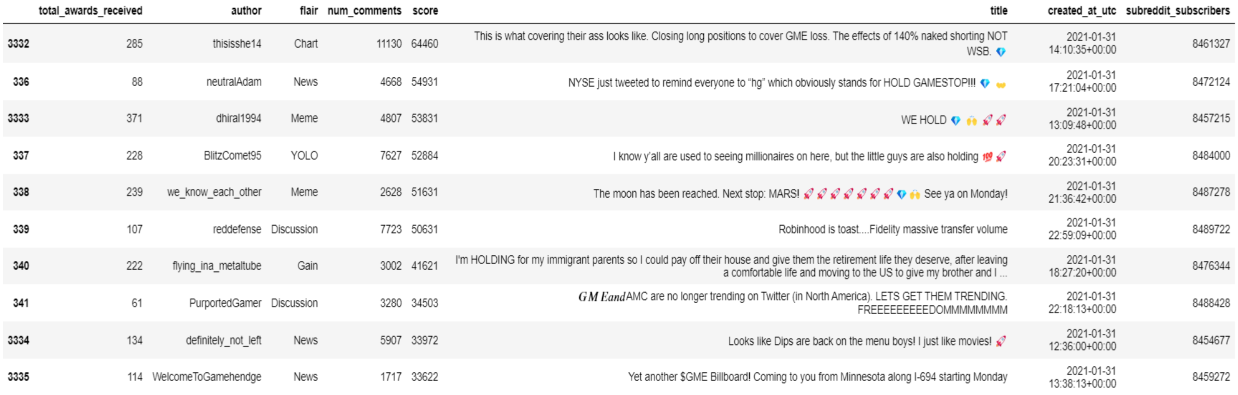
It is therefore plausible that the fluctuations in the share price of a near-bankrupt company are linked to fluctuations in the number of tweets related to the topic: many tweets will be recordings of the events that occurred.

In general, it can be seen from the infographics, how everything was born from the subreddit r/WallStreetBets, as the facts tell, and subsequently, how the attention to the topic has developed, in a more extensive way, on Twitter.

QUERIES

It is now interesting to question the data collected, manipulated and arranged in order to extrapolate information that deepens the previous analysis.

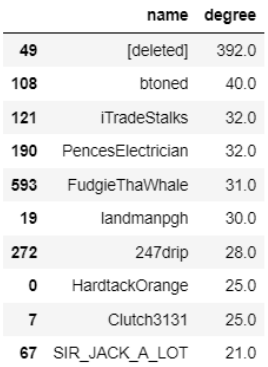
You want to view the 10 submissions that, in the time frame considered, had the highest score, therefore the posts with the greatest trend.



	total_awards_received	author	flair	num_comments	score	title	created_at_utc	subreddit_subscribers
3332	285	thisisshe14	Chart	11130	64460	This is what covering their ass looks like. Closing long positions to cover GME loss. The effects of 140% naked shorting NOT WSB. 💎	2021-01-31 14:10:35+00:00	8461327
336	88	neutralAdam	News	4668	54931	NYSE just tweeted to remind everyone to “hg” which obviously stands for HOLD GAMESTOP!!! 💎 👐	2021-01-31 17:21:04+00:00	8472124
3333	371	dhiral1994	Meme	4807	53831	WE HOLD 💎 🙌 🚀 🚀	2021-01-31 13:09:48+00:00	8457215
337	228	BlitzComet95	YOLO	7627	52884	I know y’all are used to seeing millionaires on here, but the little guys are also holding 💯 🚀	2021-01-31 20:23:31+00:00	8484000
338	239	we_know_each_other	Meme	2628	51631	The moon has been reached. Next stop: MARS! 🚀 🚀 🚀 🚀 🚀 🚀 🚀 💎 🙌 See ya on Monday!	2021-01-31 21:36:42+00:00	8487278
339	107	reddefense	Discussion	7723	50631	Robinhood is toast....Fidelity massive transfer volume	2021-01-31 22:59:09+00:00	8489722
340	222	flying_ina_metaltube	Gain	3002	41621	I'm HOLDING for my immigrant parents so I could pay off their house and give them the retirement life they deserve, after leaving a comfortable life and moving to the US to give my brother and I ...	2021-01-31 18:27:20+00:00	8476344
341	61	PurportedGamer	Discussion	3280	34503	*GM Eand*AMC are no longer trending on Twitter (in North America). LETS GET THEM TRENDING. FREEEEEEEEEDOMMMMMMMM	2021-01-31 22:18:13+00:00	8488428
3334	134	definitely_not_left	News	5907	33972	Looks like Dips are back on the menu boys! I just like movies! 🚀	2021-01-31 12:36:00+00:00	8454677
3335	114	WelcomeToGamehendge	News	1717	33622	Yet another $GME Billboard! Coming to you from Minnesota along I-694 starting Monday	2021-01-31 13:38:13+00:00	8459272

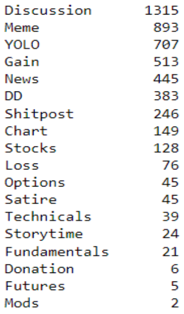
It can be seen that many submissions refer to the diamond hand, i.e., the custody of the GME shares purchased so that the price rises ’*to the moon’*.

Another interesting information is the ranking of the 10 submission authors with greater degree centrality, to display the users who have had the most influence on the social network. Below you can see the username and the degree of centrality. The user ’[deleted]’ at the top of the list is assumed to refer to all deleted submissions.



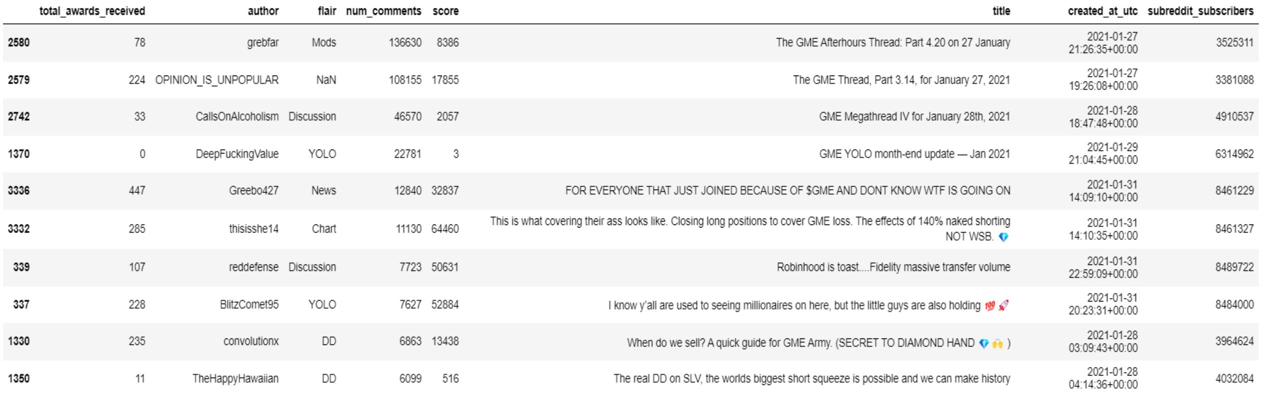
	name	degree
49	[deleted]	392.0
108	btoned	40.0
121	iTradeStalks	32.0
190	PencesElectrician	32.0
593	FudgieThaWhale	31.0
19	landmanpgh	30.0
272	247drip	28.0
0	HardtackOrange	25.0
7	Clutch3131	25.0
67	SIR_JACK_A_LOT	21.0

The distribution of the submission labels is shown below: the main ones are Discussion with 1315 submissions, Meme with 893, Yolo with 707 and Gain 513. They represent two thirds of the distribution.



```
Discussion      1315
Meme             893
YOLO             707
Gain             513
News             445
DD               383
Shitpost         246
Chart            149
Stocks           128
Loss              76
Options           45
Satire            45
Technicals        39
Storytime         24
Fundamentals      21
Donation           6
Futures            5
Mods               2
```

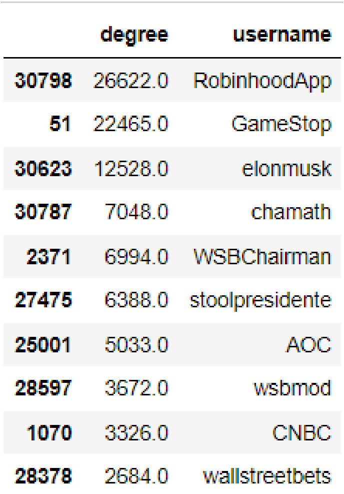
It is also useful to identify those submissions that have generated the most discussions, by extrapolating those posts that have had the most comments.



	total_awards_received	author	flair	num_comments	score	title	created_at_utc	subreddit_subscribers
2580	78	grebfar	Mods	136630	8386	The GME Afterhours Thread: Part 4.20 on 27 January	2021-01-27 21:26:35+00:00	3525311
2579	224	OPINION_IS_UNPOPULAR	NaN	108155	17855	The GME Thread, Part 3.14, for January 27, 2021	2021-01-27 19:26:08+00:00	3381088
2742	33	CallsOnAlcoholism	Discussion	46570	2057	GME Megathread IV for January 28th, 2021	2021-01-28 18:47:48+00:00	4910537
1370	0	DeepFuckingValue	YOLO	22781	3	GME YOLO month-end update — Jan 2021	2021-01-29 21:04:45+00:00	6314962
3336	447	Greebo427	News	12840	32837	FOR EVERYONE THAT JUST JOINED BECAUSE OF $GME AND DONT KNOW WTF IS GOING ON	2021-01-31 14:09:10+00:00	8461229
3332	285	thisisshe14	Chart	11130	64460	This is what covering their ass looks like. Closing long positions to cover GME loss. The effects of 140% naked shorting NOT WSB. 💎	2021-01-31 14:10:35+00:00	8461327
339	107	reddefense	Discussion	7723	50631	Robinhood is toast....Fidelity massive transfer volume	2021-01-31 22:59:09+00:00	8489722
337	228	BlitzComet95	YOLO	7627	52884	I know y'all are used to seeing millionaires on here, but the little guys are also holding 💯 🚀	2021-01-31 20:23:31+00:00	8484000
1330	235	convolutionx	DD	6863	13438	When do we sell? A quick guide for GME Army. (SECRET TO DIAMOND HAND 💎 🙌)	2021-01-28 03:09:43+00:00	3964624
1350	11	TheHappyHawaiian	DD	6099	516	The real DD on SLV, the worlds biggest short squeeze is possible and we can make history	2021-01-28 04:14:36+00:00	4032084

Having the number of comments received, it would also be interesting to evaluate whether a submission was made by a bot or by a real user. By comparing this figure with the period of activity it is in fact possible to determine the average number of comments per day, which could be suspicious.

As for the tweets, the question arose who were the most influential users in the analyzed issue. Usernames with greater degree centrality are extrapolated. It is possible to identify well-known names that have significantly influenced this event.

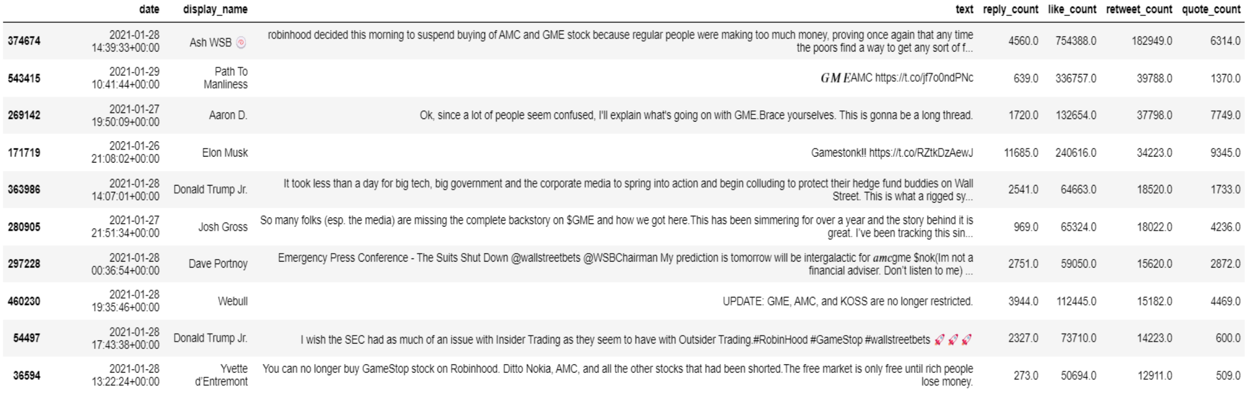


	degree	username
30798	26622.0	RobinhoodApp
51	22465.0	GameStop
30623	12528.0	elonmusk
30787	7048.0	chamath
2371	6994.0	WSBChairman
27475	6388.0	stoolpresidente
25001	5033.0	AOC
28597	3672.0	wsbmod
1070	3326.0	CNBC
28378	2684.0	wallstreetbets

RobinhoodApp is the most important user. It refers to the popular trading platform used by retail investors to buy GME shares. It played an important role during the ’GameStop saga’ by limiting transactions for the stock in question.

Other notable names are Elon Musk, who kicked off Twitter discussions about GameStop and the r/WallStreetBets subreddit with a single tweet, and high-profile venture capitalist Chamath, who via Twitter, claims to have invested in the video game company.

In addition to users, it is interesting to see the tweets that have gotten the most retweets. Among these, Elon’s famous tweet stands out Musk and the two comments made by Donald Trump Jr, who criticized Robinhood brokerage for restricting trading, and the big government and corporate media took action to protect their hedge fund friends on Wall Street, calling the rigged system. Both characters therefore take sides in favor of the small investors of the Reddit community.



	date	display_name	text	reply_count	like_count	retweet_count	quote_count
374674	2021-01-28 14:39:33+00:00	Ash WSB	robinhood decided this morning to suspend buying of AMC and GME stock because regular people were making too much money, proving once again that any time the poors find a way to get any sort of f...	4560.0	754388.0	182949.0	6314.0
543415	2021-01-29 10:41:44+00:00	Path To Manliness	*GME*AMC https://t.co/jf7o0ndPNc	639.0	336757.0	39788.0	1370.0
269142	2021-01-27 19:50:09+00:00	Aaron D.	Ok, since a lot of people seem confused, I'll explain what's going on with GME.Brace yourselves. This is gonna be a long thread.	1720.0	132654.0	37798.0	7749.0
171719	2021-01-26 21:08:02+00:00	Elon Musk	Gamestonk!! https://t.co/RZtkDzAewJ	11685.0	240616.0	34223.0	9345.0
363986	2021-01-28 14:07:01+00:00	Donald Trump Jr.	It took less than a day for big tech, big government and the corporate media to spring into action and begin colluding to protect their hedge fund buddies on Wall Street. This is what a rigged sy...	2541.0	64663.0	18520.0	1733.0
280905	2021-01-27 21:51:34+00:00	Josh Gross	So many folks (esp. the media) are missing the complete backstory on $GME and how we got here.This has been simmering for over a year and the story behind it is great. I've been tracking this sin...	969.0	65324.0	18022.0	4236.0
297228	2021-01-28 00:36:54+00:00	Dave Portnoy	Emergency Press Conference - The Suits Shut Down @wallstreetbets @WSBChairman My prediction is tomorrow will be intergalactic for *amc*gme $nok(Im not a financial adviser. Don't listen to me) ...	2751.0	59050.0	15620.0	2872.0
460230	2021-01-28 19:35:46+00:00	Webull	UPDATE: GME, AMC, and KOSS are no longer restricted.	3944.0	112445.0	15182.0	4469.0
54497	2021-01-28 17:43:38+00:00	Donald Trump Jr.	I wish the SEC had as much of an issue with Insider Trading as they seem to have with Outsider Trading.#RobinHood #GameStop #wallstreetbets 🚀🚀🚀	2327.0	73710.0	14223.0	600.0
36594	2021-01-28 13:22:24+00:00	Yvette d'Entremont	You can no longer buy GameStop stock on Robinhood. Ditto Nokia, AMC, and all the other stocks that had been shorted.The free market is only free until rich people lose money.	273.0	50694.0	12911.0	509.0

CONCLUSIONS and FUTURE DEVELOPMENTS

From the analysis carried out, it was quite clear how the social networks conducted the campaign that triggered the short squeeze of the GME stock. In particular, it all started on Reddit, after which the increased volume of message exchanges continued on Twitter.
The involvement of well-known names in the world of finance and the names of influential people is evident.
We can realize that the stock trend can be determined through unconventional channels, as can the world of social media.

A further investigation could concern the analysis of the text of the messages posted on the two social platforms, through the identification of bots and a sentiment analysis.
Furthermore, clustering algorithms could be used to identify possible groups of users with common behaviors and relationships.

REFERENCES

https://en.wikipedia.org/wiki/GameStop_short_squeeze

https://abcnews.go.com/Business/GameStop-timeline-closer-saga-upended-wall-street/story?id=75617315

https://abcnews.go.com/Business/reddit-users-GameStop-stock-soaring-upending-market/story?id=75513249

Snscrape
https://github.com/JustAnotherArchivist/snscrape

Pushshift
https://github.com/pushshift/api

Neo4J
https://neo4j.com/