An analysis of WSB with distributed computing

Team Diamond Hands - Kenny, Steven, Callen

Project Goal

🚀

🚀

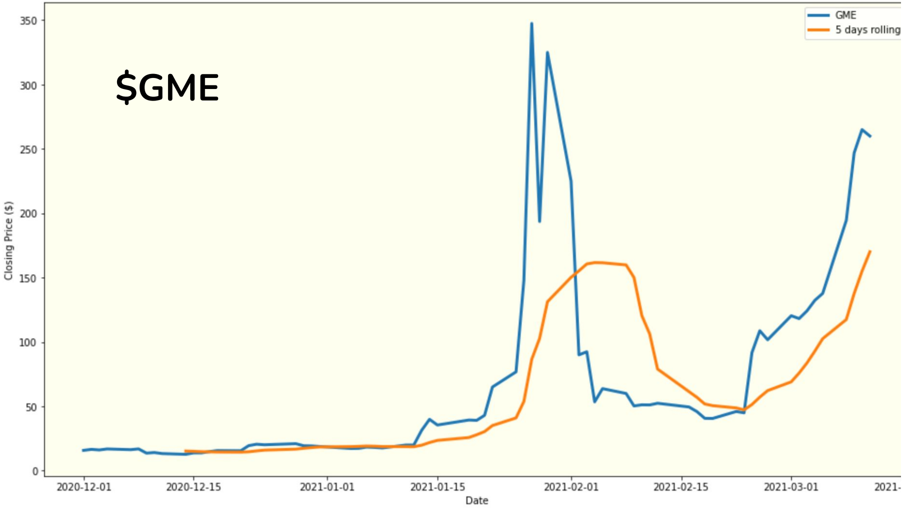
- Use the Twitter and/or Reddit API to scrape comments on these services for discussion on stocks (specifically volatile stocks ‘meme’ stocks such as GME and AMC).
- Perform ETL processes in a distributed manner using Spark
- Basic sentiment analysis on what people think of these stocks
  - Examine discussion over time
  - Cross compare discussion to the trends of stock market
  - Analyze the volume of interest in these stocks

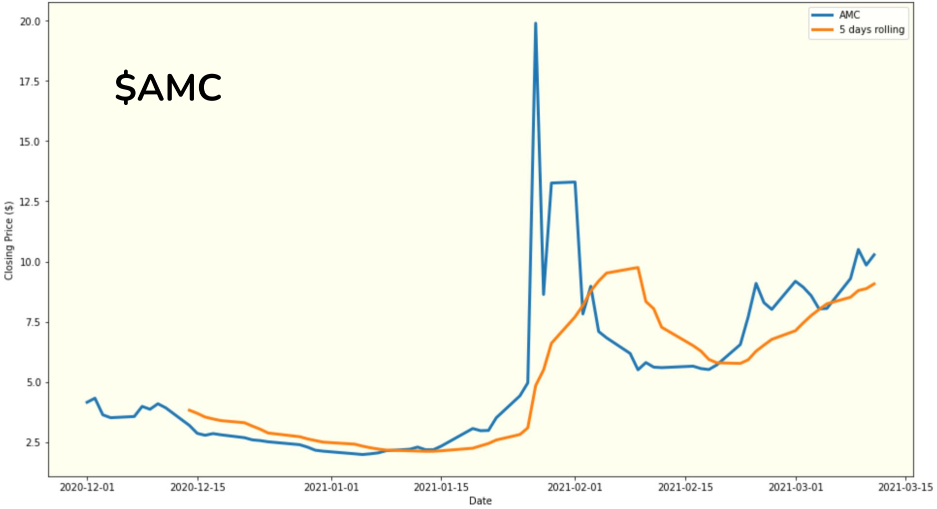
💎 💎 💎

The Behavior over time…

Key Dates:

- GME & AMC ATH both on Jan. 27th 2021 at $347 & 19.90 respectively
- Buying restrictions: Jan. 26th - Jan. 29th 2021
- Congressional hearing: Feb. 18th 2021



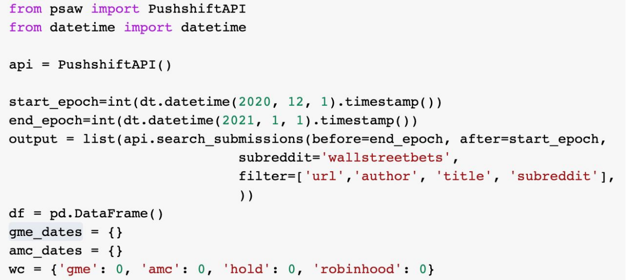


Schedule

Task Name	W5	W6	W7	W8	W9	W10	Deadline
Initial Research	■	■					2/11
Twitter/Reddit/ Finance API Investigation	■	■					2/11
System Design		■					2/11
Continuous Research		■	■				2/18
Data preprocessing			■				2/18
Product development		■	■	■	■		3/4
Product refinement					■	■	3/11
User Testing						■	3/11
Documentation						■	3/11

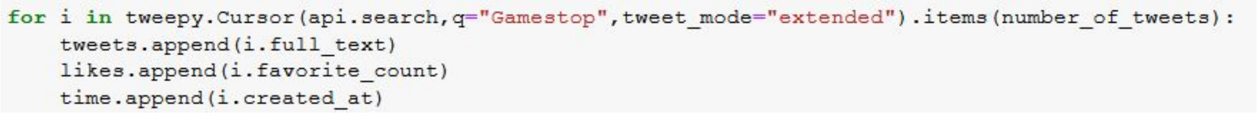
Research

Reddit API -



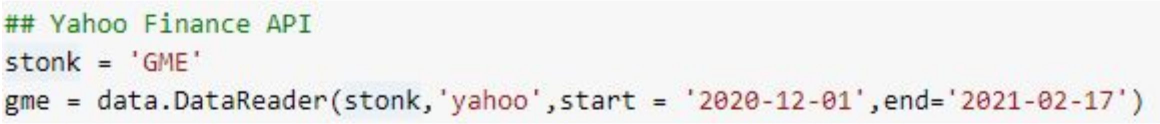
```python
from psaw import PushshiftAPI
from datetime import datetime

api = PushshiftAPI()

start_epoch=int(dt.datetime(2020, 12, 1).timestamp())
end_epoch=int(dt.datetime(2021, 1, 1).timestamp())
output = list(api.search_submissions(before=end_epoch, after=start_epoch,
                            subreddit='wallstreetbets',
                            filter=['url','author', 'title', 'subreddit'],
                            ))
df = pd.DataFrame()
gme_dates = {}
amc_dates = {}
wc = {'gme': 0, 'amc': 0, 'hold': 0, 'robinhood': 0}
```

Twitter API -



```python
for i in tweepy.Cursor(api.search,q="Gamestop",tweet_mode="extended").items(number_of_tweets):
    tweets.append(i.full_text)
    likes.append(i.favorite_count)
    time.append(i.created_at)
```

Yahoo Finance API -

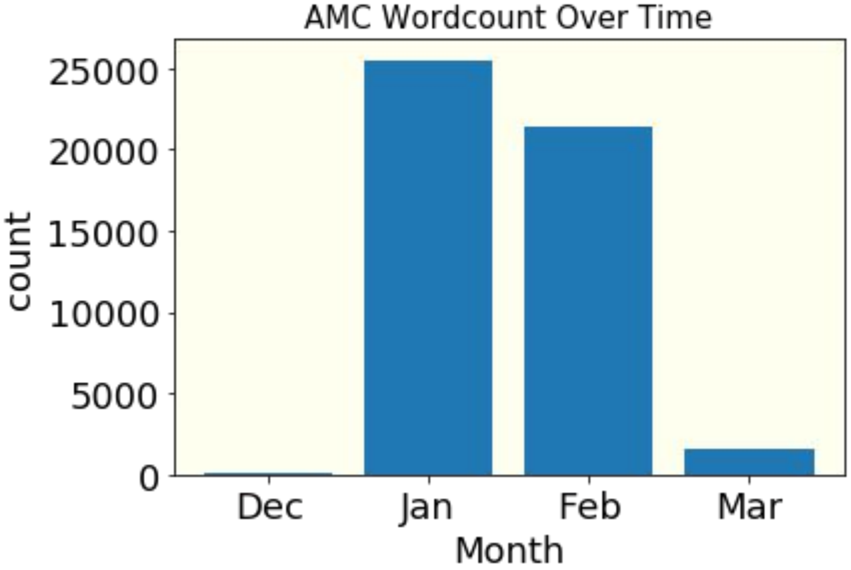


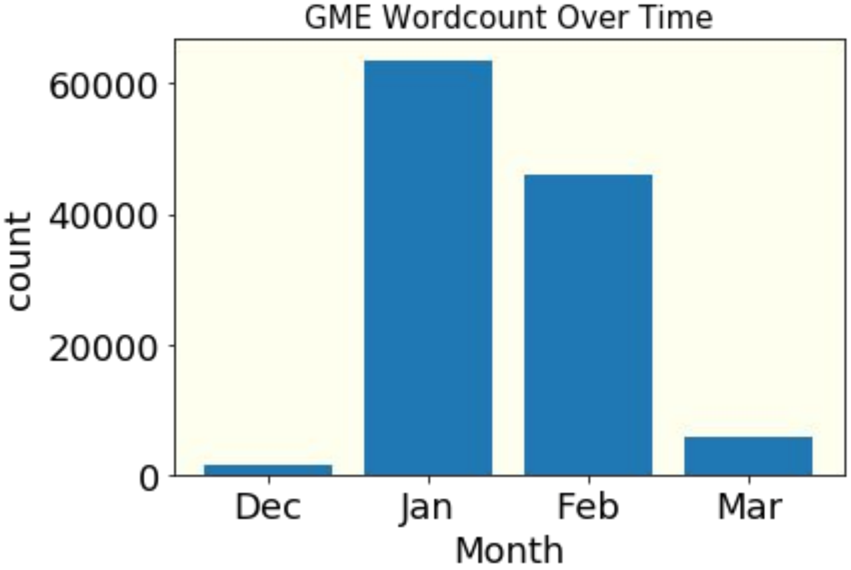
```python
## Yahoo Finance API
stonk = 'GME'
gme = data.DataReader(stonk,'yahoo',start = '2020-12-01',end='2021-02-17')
```

Scraping & Processing WSB Posts

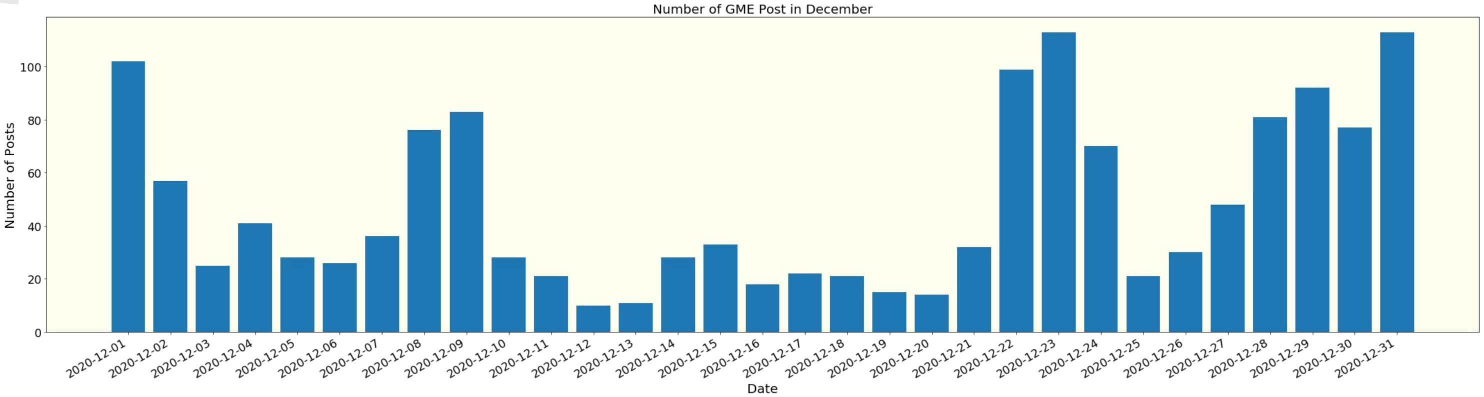
- **PSAW API -> PySpark -> Pandas -> Matplotlib**
- PSAW
  - Scraped all posts from r/WallStreetBets from December 2020 - March 2021
    - Over 150MB of data
  - Data included post title, date, links, author
- PySpark
  - Used Spark SQL, map & reduce to aggregate and extract features from data
- Converted Spark DF to Pandas DF to visualize data

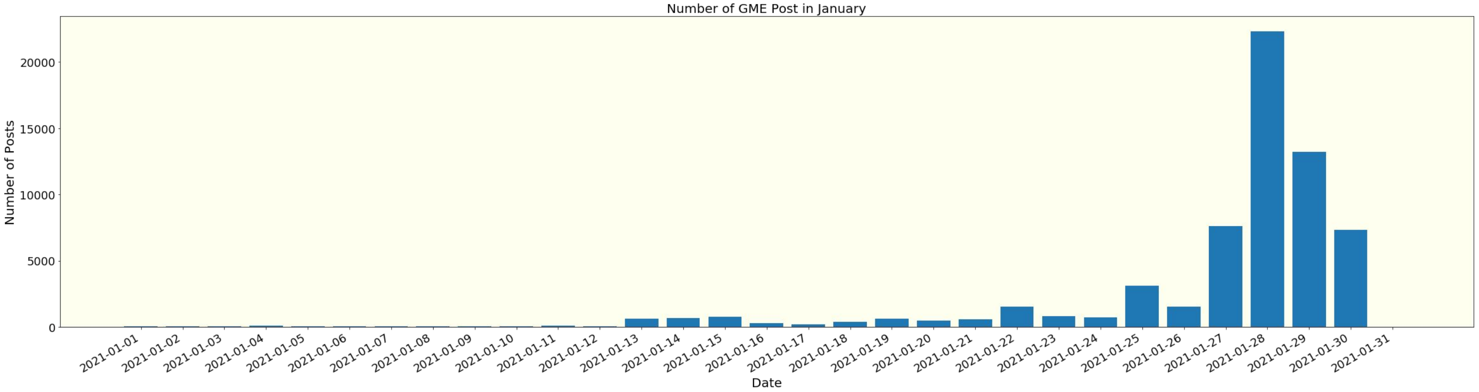
AMC & GME Word Count



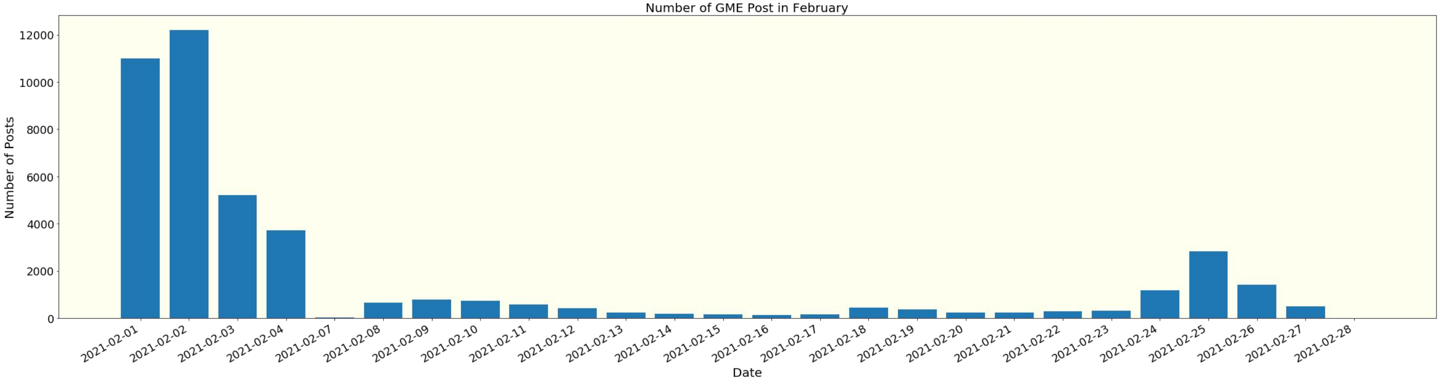


GME Post By Month

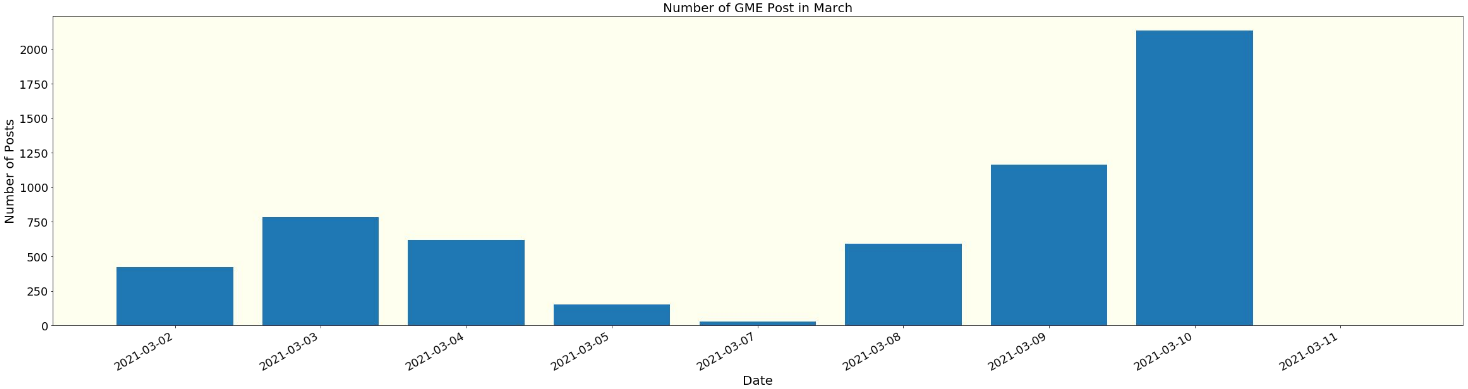




GME Post By Month

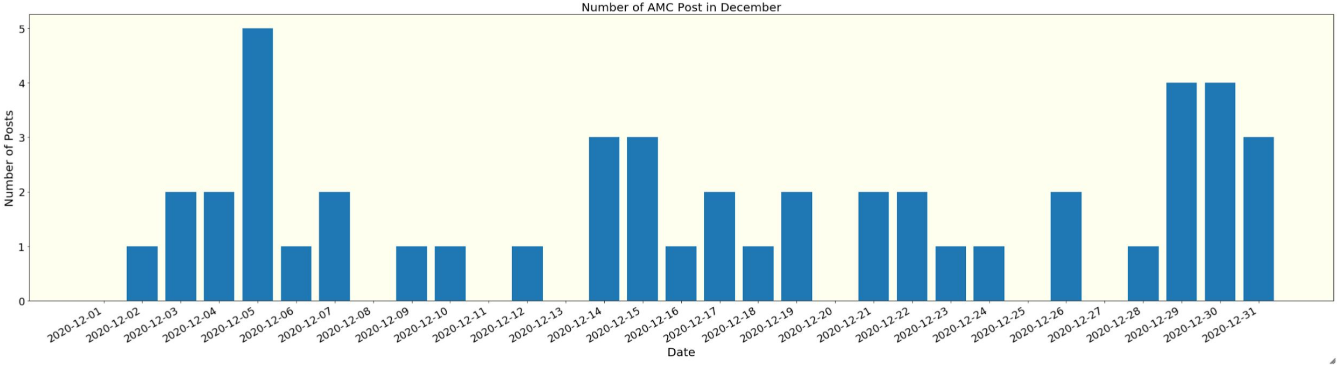


Number of GME Post in February

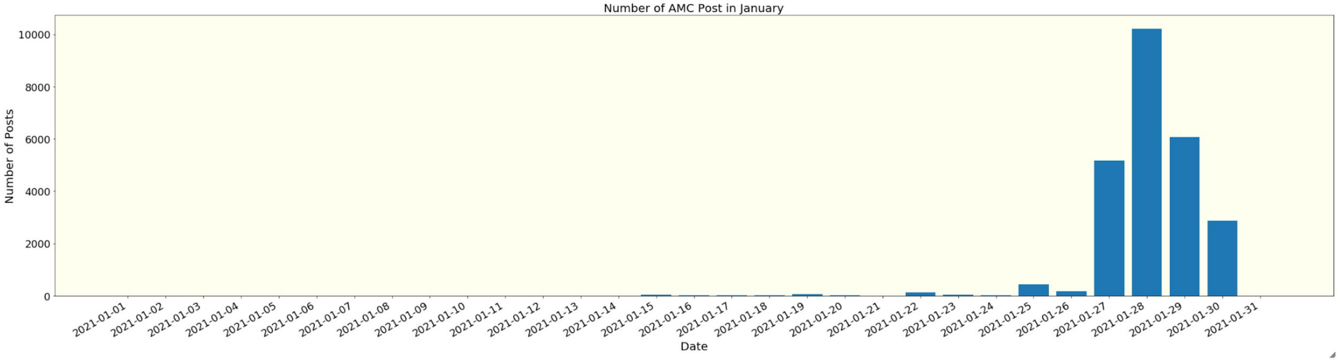


Number of GME Post in March

AMC Post By Month

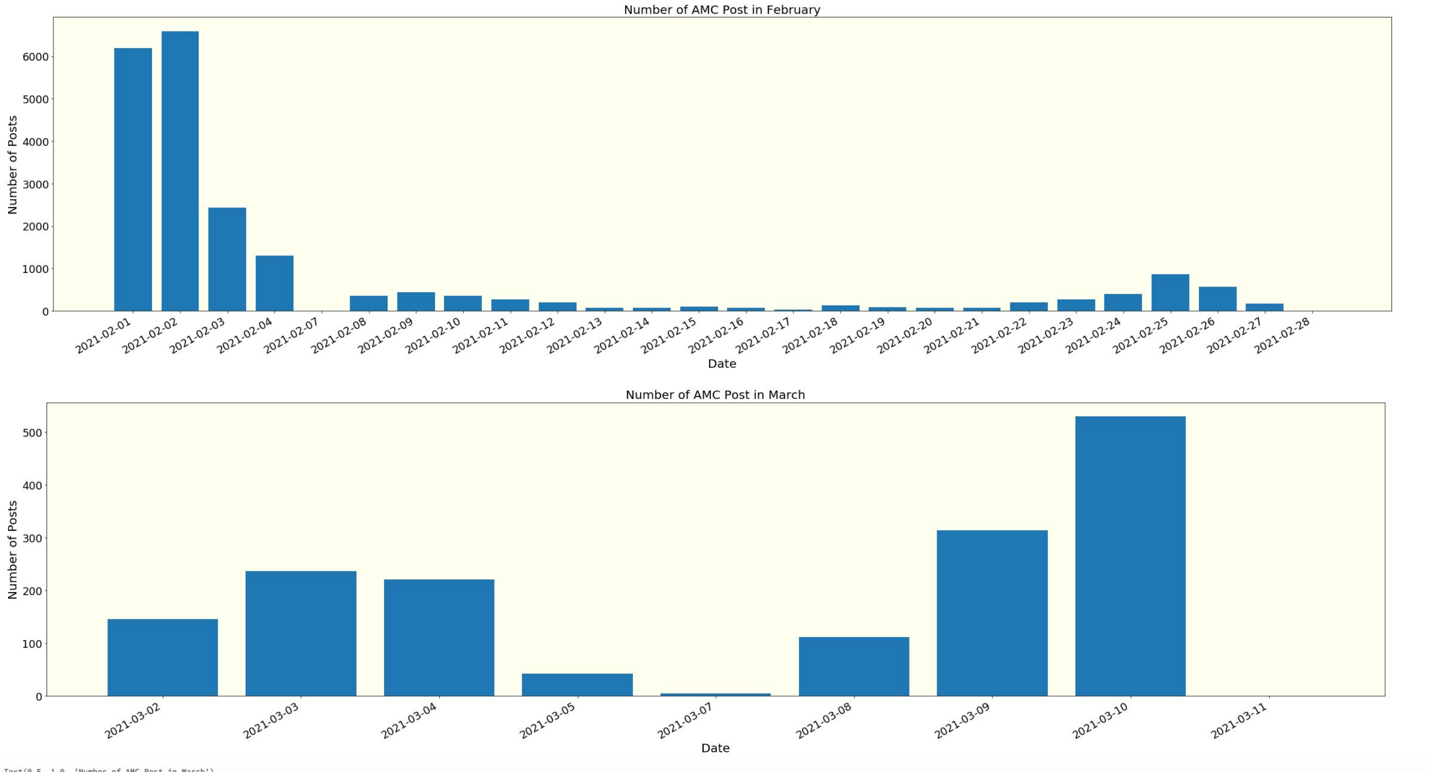


Number of AMC Post in December



Number of AMC Post in January

AMC Post By Month

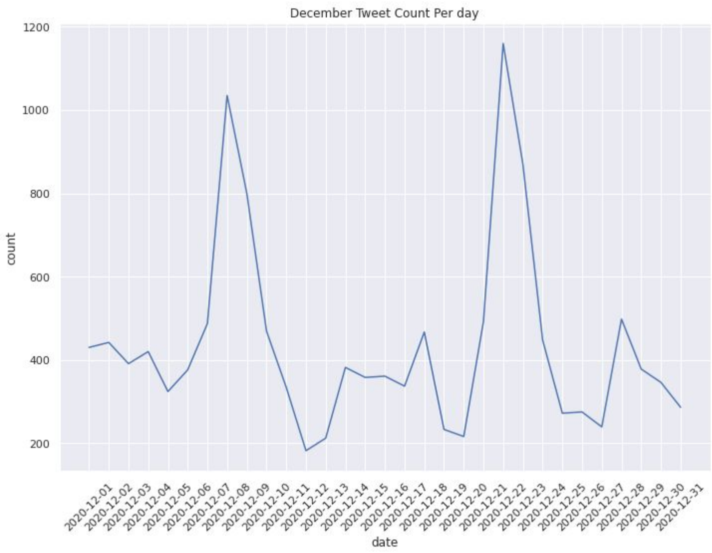


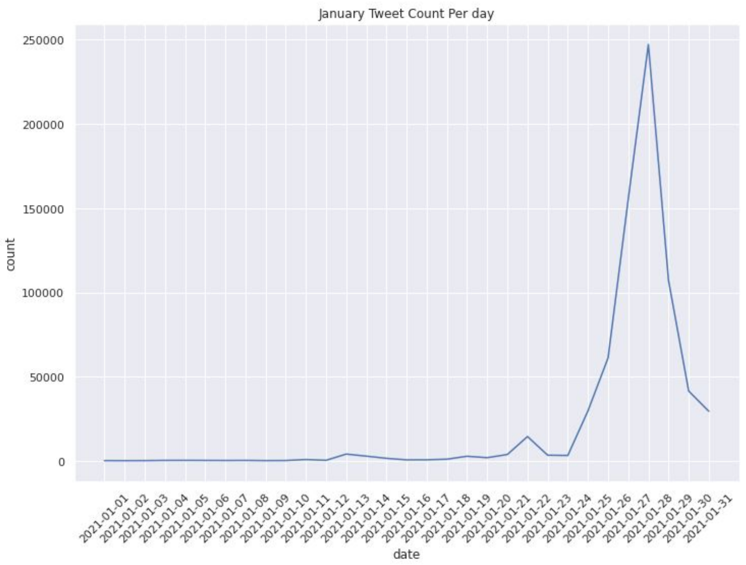
```
Text(0.5, 1.0, 'Number of AMC Post in March')
```

Scraping & Processing Tweets

- **Twint API -> PySpark -> NLTK -> Pandas ->Matplotlib**
- Twint API
  - Scraped all tweets containing “GME” for each month (Dec 2020 - Mar 2021)
  - Up to 700,000 tweets per month
  - Pulled features such as text, like count, retweet count, language, etc.
- PySpark
  - Used Spark SQL queries to select specific variables from large datasets
  - Used map & filter functions to process data in a distributed fashion
    - Ex. Removing stop words and irrelevant characters w/ NTLK
- Converted Spark RDD’s to Pandas Data Frames for plotting/analysis

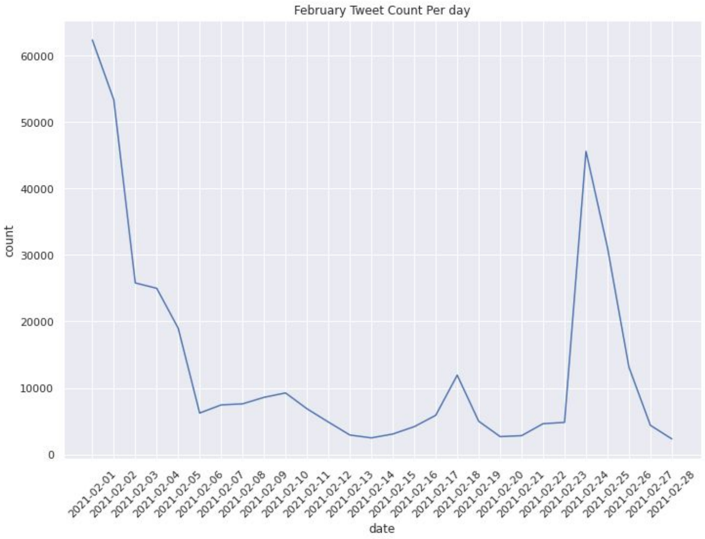
Number of Tweets Over Time

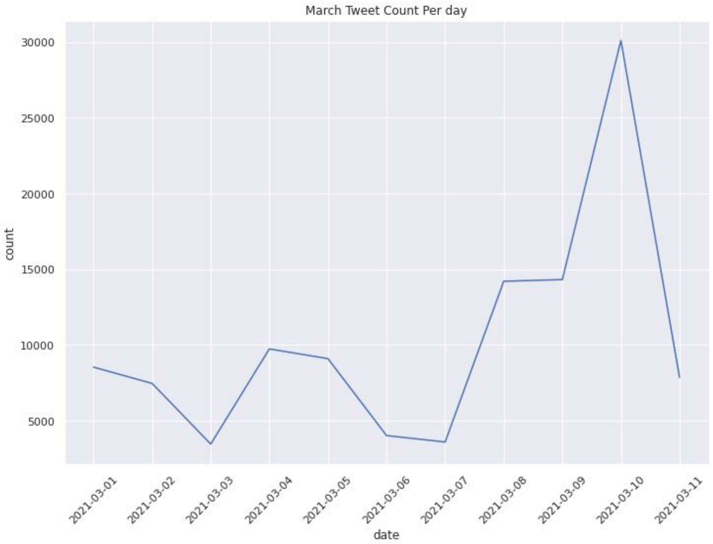




- Very little engagement in December (around 200-1000 tweets per day). Massive increase of 250,000 tweets during initial GME spike and temporary buying restrictions (1/26 - 1/29).

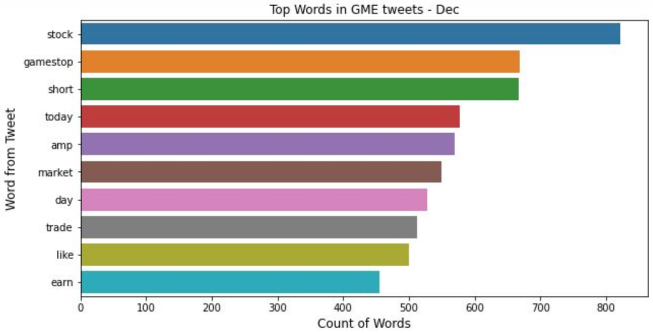
Number of Tweets Over Time

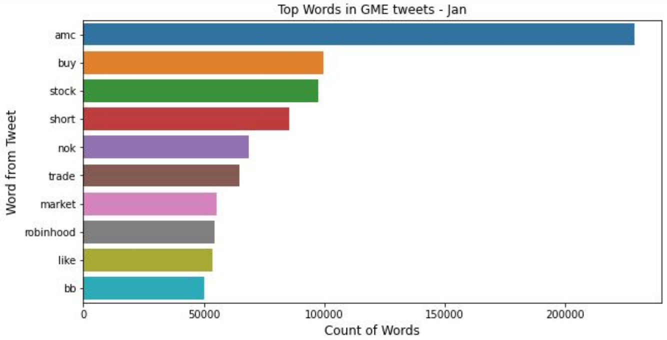

February Tweet Count Per day

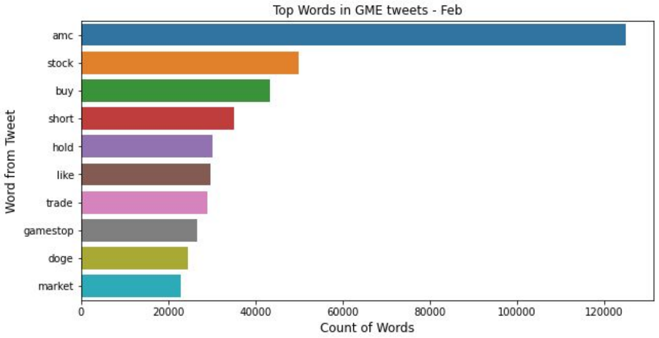

March Tweet Count Per day

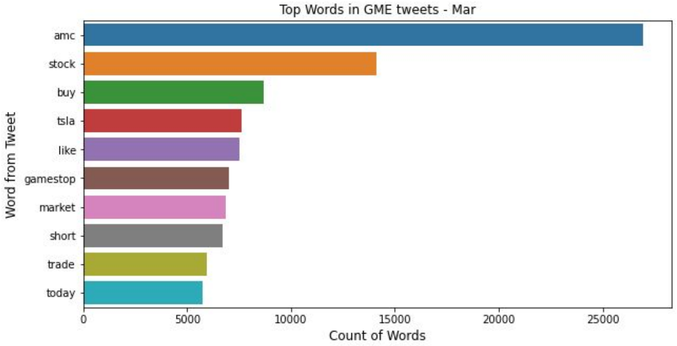
- Downward trend of tweets after first spike in GME. Engagement returns during Congressional Hearing (2-18) and second rise(2-23 ->3-10).

Top Words from Tweets By Month


Top Words in GME tweets - Dec


Top Words in GME tweets - Jan


Top Words in GME tweets - Feb


Top Words in GME tweets - Mar

What Can We Gather From Our Results?

- Distributed computing helps a ton in processing data
  - Much more efficient in comparison to for-loops
- GME and AMC both jumped to insane popularity in January and early February
  - Popularity is not as high, but there looks to be new interest in March
- Tweets and Reddit posts increased engagement when GME stock had spiked or had been restricted.
  - Specific keywords may have indicated interest in buying, selling, or holding stocks

Next Steps…

- Stream data through Spark for real-time analysis
- Retrieve data from different stocks and media platforms
- Develop Machine Learning /NLP Models to predict stock trends

Thank you!