

Can Reddit Sentiment Affect the Stock Market?

Group Number 2

Derek Bobbitt, Macy Broderick, Matthew Clarke, and
Claire Plourde

December 22nd, 2023

Table of Contents

Executive Summary	3
Introduction	3
Dataset Selection	4
Dataset Choice	4
Duplication Penalty Acknowledgment	4
Data Management	5
Process Detailing	5
Transformation and Integration	5
Data Exploration and Analysis	6
Description and Trends	6
Text Analysis	7
Text Preprocessing	7
Sentiment Analysis	7
Interpretation	8
Statistical Modeling	9
Model Application	9
Data Visualization	10
Visualization Development	10
Visualization Interpretation	15
Conclusion	16
References	17
Appendix	17

Executive Summary

Our team conducted an in-depth analysis of the potential correlations and relationships between Reddit comments pertaining to Apple's stock price (\$AAPL) and the performance of the stock in terms of price. The current investment landscape has adapted with hedge funds introducing new ways to trade based on social media sentiment utilizing machine learning in what some investors call opinion mining. In an effort to understand the syntax and performance of this method, we leveraged R to allow for sentiment and statistical analysis that further enhanced our group's perspective on the applicability of these algorithms. Through these methods, we were able to see trends suggesting that any large changes in Apple's stock price led to an increase in activity with Reddit comments and posts relating to the company. Due to Apple's success over the timeframe associated with the Reddit dataset, further research on other companies' stock price and public sentiment would need to be done in order to provide recommendations.

Even without being able to provide evidence of the performance of sentiment gathering algorithms, trends were apparent to indicate that Apple's performance over the five-year period led to an increase in "mentions" either in comments or in posts. Through this finding, we concluded that over this time period, Reddit users were becoming more interested in financial markets and took to public forum boards to share their opinions on certain stocks. Using word clouds, we were able to build a narrative around the common positive words and the negative words to see if people were just hoping to generate hype around the stock or provide criticism to improve one's short position.

Overall, we found there was a slight relationship between number of posts, sentiment of these posts, and the AAPL stock price. However, it is not clear if this is the most profitable way to predict stock price as Reddit comments and posts increased with any stock activity, growth or decline.

Introduction

The GameStop Saga in January 2021 was the first time social media was truly brought to mainstream attention as a form of influence in financial markets. Although it was common in day and quantitative trading communities to scrape social media data as part of their analysis or data for training trading bots, analyzing social sentiment now became a crucial step across most trading strategies, even in corporate finance. The idea that a group can organize through social media, invest in sync or at the very least in tandem, had never been seen before at the scale of the r/wallstreetbets community. Now that this is a reality seen in the GameStop Saga, we've seen this trend grow greatly. However, can we really say it is worth it to look at any one social media platform and find a statistically significant correlation when utilizing sentiment analysis? In this research project, we aim to explore the relationship between Reddit sentiment and stock performance.

For the scope of our research, we chose to investigate if sentiment could really move some of the world's largest companies or if GME was the exception, not the rule. Accordingly,

we chose to look at Apple, the largest company in the world. This provided this project with a lot of coverage and an abundance of data.

Dataset Selection

Dataset Choice

The dataset we selected was labeled “Five Years of AAPL on Reddit”. This dataset contains Reddit comments and posts from November 2016 to October 2021 that mention “AAPL”, Apple’s stock ticker. The dataset is split into two CSV files, posts and comments. In both files, fields like created date, subreddit name, body of the post or comment, sentiment, and score, were provided. This collection of data allowed the potential for text analysis of the posts and comments, including sentiment and topic modeling, analysis across subreddits, and time series analysis.

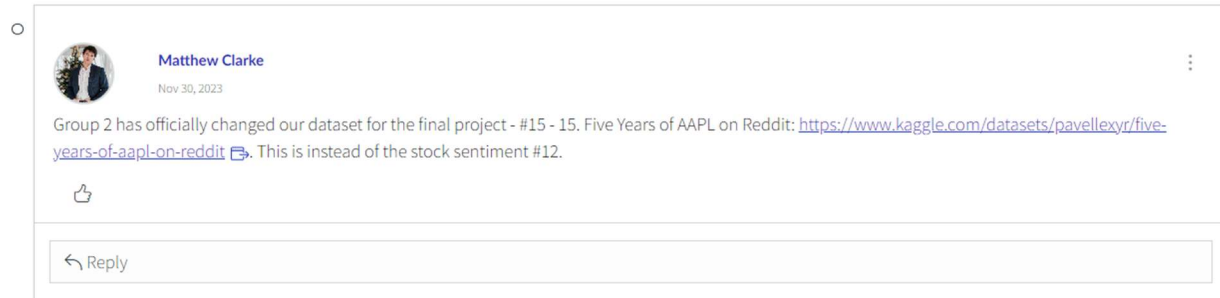
Comments Data Dictionary	
Name	Description
type	Type of the datapoint
id	Unique Base-36 ID of the comment
subreddit.id	Unique Base-36 ID of the comment's subreddit
subreddit.name	Human-readable name of the comment's subreddit
subreddit.nsfw	Is the comment's subreddit NSFW?
created_utc	Timestamp of the comment's creation
permalink	Permalink to the comment on Reddit
body	Comment's body text
sentiment	Analyzed sentiment for the comment on score from -1 to 1
score	Comment's score

Posts Data Dictionary	
Name	Description
type	Type of the datapoint
id	Unique Base-36 ID of the post
subreddit.id	Unique Base-26 ID of the post's subreddit
subreddit.name	Human-readable name of the post's subreddit
subreddit.nsfw	Is the post's subreddit Not Safe For Work?
created_utc	Timestamp of the post's creation
permalink	Permalink to the post on Reddit
domain	Base url of the below url field
url	Link to any content posted within the post (youtube video, article, etc)
selftext	The body of the post
title	The title of the post
score	Post's Score

In addition to the Reddit dataset, we used a dataset labeled “AAPL Daily Stock Performance” from Yahoo Finance to enhance our existing data. The combination of the two datasets allowed us to analyze the relationship between Reddit posts and comments and Apple’s stock price over time. The data contained in these datasets includes values that were useful for sentiment and time-based analysis. Because we had the body of the posts and comments, we were also able to analyze the relationship between stock and sentiment of posts and comments over time.

Duplication Penalty Acknowledgment

We confirm that we have verified no other group is using the same dataset. None of the other groups stated on the discussion board they are using this dataset and none of the presentations in class on December 14th, 2023 used this dataset or covered this topic.



Confirmation of Dataset Claim

Data Management

Process Detailing

The Reddit dataset was already moderately clean. Our data cleaning was completed in Excel. The created datetime was displayed in UTC, a date format that displays the number of seconds since January 1st, 1970. It is fairly unreadable and difficult to analyze so we converted this to a datetime format with the formula, $\{\text{UTC Date}\}/(60*60*24)+\text{"1/1/1970"}$. The column NSFW, which indicates if a subreddit is Not Safe For Work, was displayed as True or False, which can be difficult to analyze. This was changed to 1 for True and 0 for False using find and replace. A few subreddit names were missing or displayed as "???". Luckily, the link to each post and comment was included in the dataset. We visited these links, read the name of the subreddit, and replaced the blank or "???" subreddit name with the correct name. Sentiment and score had been read in as text fields, these were converted to numeric fields. There were quite a few missing sentiment values that we replaced with 0, a neutral value.

The Apple stock dataset was also moderately clean when received. The High and Low columns were removed as they were out of scope of this project. A Daily % Change field was added using the Open and Close columns.

Once the data was clean, we loaded it into R to check for any missing data and duplicates. There were no duplicates and all missing data had been replaced. The body column was converted to lowercase for consistent text analysis. A length of comment field was added to aid in analysis.

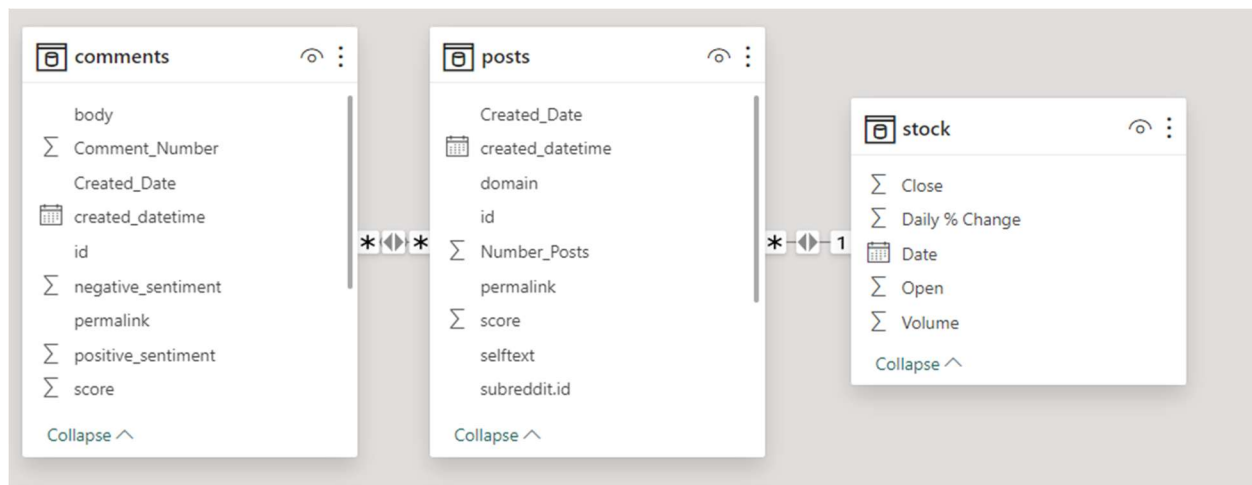
```
#add length of comment
comments = comments %>%
  mutate(comment_length = sapply(strsplit(body, " "), length))
```

Adding Comment Length Field in R

Transformation and Integration

The data, both Reddit and Stock, were already normalized when received. Joining the data was applied in PowerBI. All three datasets were joined on the date field, a calculated field, created_date for comments and posts and Date for stock. The created_datetime field in posts and

comments was not used because we would not be analyzing as granular as every hour or every minute. We also did not obtain stock data for every minute or every hour, just daily.



Model View in PowerBI

We kept comments and posts as separate datasets even though they could have been joined with a union. Joining the two datasets would not have benefited our analysis in any way and would have only made a larger dataset, which would have potentially slowed down any application we used to analyze.

Data Exploration and Analysis

Description and Trends

Our dataset contained 297,533 comments and 15,483 posts across 2,669 unique communities on Reddit, also known as subreddits. The subreddit with the most posts was r/wallstreetbets with 3,730 posts. The subsequent subreddits with the most posts were r/stocks, r/optionmillionaires, r/newsbotmarket, and r/forexhome, all between 600-800 posts. It is clear r/wallstreetbets is the most popular subreddit to discuss AAPL and stocks in general. Knowing which subreddits the posts and comments came from is important because different subreddit may have different opinions on AAPL and stocks in general. If specific subreddit that have a relationship with the AAPL stock can be identified, this would help greatly in our analysis of the relationship overall.

The average sentiment for comments, on a scale from -1 to 1, was 0.15. This is slightly more positive than neutral. The average score for comments was 5.19 and the average score for posts was 27.36. This tells us that posts are more likely to be upvoted than comments, probably due to the higher visibility and lower number of posts. It is good to be aware of this because there is a slight bias towards positive sentiment. While conducting our analysis, we will keep this in mind while analyzing trends.

Text Analysis

Our text data was the richest source of information in our dataset and we utilized this data in a number of different ways. We investigated the relationship between the sentiment of the Reddit comments and the performance of the Apple stock data. However, we also analyzed how other factors may influence the sentiment of the comment such as the length of the comment or the subreddit that the comment is involved in. Finally, we looked at words that were most commonly used in the positive comments compared to the negative comments to get a better understanding of what our sentiment data meant.

Text Preprocessing

The creation of our word clouds required the most text preprocessing to ensure accurate and meaningful results. We began by filtering our comments data frame into 2 separate data sets. In one, we chose only comments with scores greater than 0. This dataset held our positive comments. In the other, we chose only comments with scores less than 0. This dataset held our negative comments. The data from the "body" column of each dataset was compiled into a corpus, and all of the words were changed to lowercase using the `tolower()` tool. The reason that we changed this text to lowercase was because if some words were in different cases, they would be recorded as separate words, which could impact our results. Specifically, we were looking to create word clouds using this text data, which present the most frequently used words in a set of data. If the tool registers the same word as two different words based on their capitalization, this causes our word clouds to be inaccurate. Following this, all punctuation, numbers, and stop words were removed. In addition to the commonly accepted collection of stop words, we also removed the words "aapl", "stock", "apple", "stocks", "like", "will", and "just" as they were either related to apple stocks in general or were deemed too commonly used but insignificant to our analysis. Once this was complete, our datasets were ready to be used.

In terms of the sentiment scores, our dataset had already calculated a sentiment score for each comment that we utilized in our analysis. This data included sentiment scores ranging from -1 to 1 , 1 being the most positive and -1 being the most negative. We utilized this data in our analysis, however the sentiment of text data can be analyzed in a number of different ways. This method was reapplied when analyzing sentiment with posts, as this variable was not included in the original dataset. To compensate for this, our group utilized R to split the posts' words into tokenized variables and then apply a preset lexicon called `bing sentiment` to individually consider each word negative, positive, or neutral. By breaking down the comments and posts into individual words, the function matches the words to the lexicon and then stores the corresponding sentiment. For comments, this led to roughly 450,000 positive and about 350,000 negative sentiment, whereas for posts, the numbers were $\sim 6,000$ and $\sim 4,000$ respectively.

Sentiment Analysis

Methodology

We utilized the sentiment column within our comments data frame to analyze how the activity on Reddit, specifically the sentiment of the comments where the Apple stock ticker (AAPL) was mentioned, may be related to Apple stock performance. We conducted a time series

analysis using the sentiment and timestamp data in the comments data frame. We generated a visualization with this data to see if there were any notable trends in the data. We then added another line in this same visualization showing the trends in Apple stock performance for this same time period. In addition to the Apple stock performance data, we also looked into the significant events that took place at the times that our data seemed to diverge from its usual trends.

We also investigated the specific factors that may affect a comment's sentiment such as the subreddit that it was involved in and the comment's length. We went on to complete this analysis using a Multiple Linear Regression model, of which we will go into more detail in the subsequent section.

We then looked at what the most frequently used words in the comments with a negative sentiment versus comments with a positive sentiment. As mentioned, we created one dataset containing only positive comments and one containing only negative comments. We created two separate word clouds for each of these two datasets. The word clouds were then developed based on the most repeated words in each of these corpuses using the R code: *wordcloud()*. Within this code we set a few additional parameters such as the scale of the visual, the maximum words used to be 100, and we ensured that the words were not randomized.

Findings

Through the analysis of our text data, we found that there didn't seem to be too much of a relationship between Apple stock performance and sentiment of comments related to the Apple stocks on Reddit, however, there were some notable trends. When the Apple stock's performance began to increase substantially, the sentiment of comments did drastically jump, and began to fluctuate much more during that increase.

When looking at the relationship between sentiment and other potential influences, there does seem to be a statistically significant relationship to comment length, but there was no indication of relationship with the subreddit. This is discussed more in detail in the Modeling section.

In terms of the word cloud, there were very clear differences between the two, however there were also many similarities. "Market", "buy", "time", "money", and "think" were some frequently used words in both clouds. However, there were also some clear differences. For instance, in the negative word cloud we saw a lot of swear words, as well as words like "short", "loss", "risk" which indicate more of a negative feeling of money lost. In the positive word cloud we saw lots of words indicating growth and wealth such as "good", "right", "earnings", and "growth".

Interpretation

These results indicate to us that there are specific trends in the sentiment of reddit comments over time that are worth looking into, and there may be a relationship between sentiment and the performance of the stock. There are several significant events that occur at the same time as some of these trends that may have also had an impact on the sentiment of the

comments. Additionally, these results indicate that there is a relationship between comment length and sentiment, but no significant relationship between the sentiment and the subreddit. This introduces other possible explanations behind the sentiment of the comments, outside of our initial analysis related to the stock performance.

Lastly, our glimpse into the most frequently used words in positive versus negative comments showed differences that made sense in terms of our data, as some of the words in the negative word cloud were synonymous with loss, while some of the positive words were synonymous with growth. However, there were several words in each of the clouds which could be used in either a positive or negative context, and it is up to the lexicon to determine what the sentiment of those will be. For example, the negative cloud contained a lot of swear words, but often on social media people will use swear words to express positive sentiment. Although the lexicon assigned these comments to be either positive or negative, we would not truly understand the sentiment without reading each one individually to bring in context.

Statistical Modeling

Model Application

After interpreting the correlation matrix's results, our team aimed to apply concepts learned in DS 803, specifically the Multiple Linear Regression model, to see if the Subreddit Name had any statistically significant impact on the sentiment. The first image below shows the formula including comment length and the factor of subreddit name, which is the process of breaking up the variable into individual segments to provide a more precise relationship. The results suggest that comment length has an impact on sentiment as well as some individual subreddit names, but for the majority of the subreddits, there was little to no significance. It should be noted that the R^2 was ~ 0.153 , which means 85% of the variance is random and is not explained by these variables.

```
factor(subreddit.name)lucidmotorsinvestors -6.062e-01 5.634e-01 -1.076 0.281958
factor(subreddit.name)m1finance -1.990e-01 3.563e-01 -0.558 0.576556
factor(subreddit.name)macmini -1.624e+00 5.634e-01 -2.882 0.003956 **
factor(subreddit.name)magic leap -4.685e-01 5.634e-01 -0.832 0.405695
factor(subreddit.name)makerdao 1.855e-01 4.600e-01 0.403 0.686773
factor(subreddit.name)maticnetwork -2.378e-01 5.634e-01 -0.422 0.673014
factor(subreddit.name)maxjustrisk -1.205e+00 4.614e-01 -2.611 0.009039 **
factor(subreddit.name)me_irl -4.038e-01 5.634e-01 -0.717 0.473573
factor(subreddit.name)mechmarket -5.907e-01 5.634e-01 -1.048 0.294500
factor(subreddit.name)memeconomy -9.887e-01 5.634e-01 -1.755 0.079322 .
factor(subreddit.name)mgto 2.338e-01 5.634e-01 0.415 0.678183
factor(subreddit.name)mildlyinfuriating 1.236e-01 5.634e-01 0.219 0.826388
factor(subreddit.name)mildlyinteresting -9.157e-01 5.634e-01 -1.625 0.104134
factor(subreddit.name)militaryfinance 7.142e-02 5.634e-01 0.127 0.899132
factor(subreddit.name)millennialbets -1.564e+00 5.634e-01 -2.776 0.005520 **
[ reached getOption("max.print") -- omitted 237 rows ]
```

```
Call:
lm(formula = sentiment ~ comment_length + factor(subreddit.name),
    data = comments_data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.32716 -0.26746 -0.04279  0.36418  1.17272

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.855e-01  3.253e-01   1.800  0.071940 .
comment_length    1.037e-03  6.888e-05  15.061 < 2e-16 ***
factor(subreddit.name)4chan      8.444e-02  5.634e-01   0.150  0.880865
factor(subreddit.name)4kto1m     2.825e-01  5.634e-01   0.501  0.616074
factor(subreddit.name)5kto100kchallenge -2.845e-01  5.634e-01  -0.505  0.613641
factor(subreddit.name)aapl      -3.771e-01  3.429e-01  -1.100  0.271478
factor(subreddit.name)aaplwheel  2.192e-01  5.634e-01   0.389  0.697285
factor(subreddit.name)accounting -6.706e-01  5.634e-01  -1.190  0.234006
factor(subreddit.name)activeoptiontraders -3.490e-01  5.634e-01  -0.619  0.535675
factor(subreddit.name)algotrading -2.839e-01  3.494e-01  -0.812  0.416552
factor(subreddit.name)alpp      -1.721e-01  5.634e-01  -0.305  0.760067
factor(subreddit.name)ama       8.821e-02  4.600e-01   0.192  0.847948
factor(subreddit.name)amcstock  -4.320e-01  3.514e-01  -1.230  0.218858

Residual standard error: 0.46 on 8621 degrees of freedom
(942 observations deleted due to missingness)
Multiple R-squared:  0.1533,    Adjusted R-squared:  0.1105
F-statistic: 3.58 on 436 and 8621 DF,  p-value: < 2.2e-16
```

In the below model, we considered the impact of the factored subreddit name variable on posts' score since the Posts dataset did not have a sentiment variable provided. There were zero statistically significant subreddit names, with an incredibly low R^2 to suggest that a significant portion of the variability could not be explained. This indicates that the subreddit name is not a factor in how a post performs when it comes to score. All subreddits exhibit similar behavior when it comes to upvoting and downvoting posts, which is helpful to know if looking to incorporate post score into any analysis.

```
Call:
lm(formula = score ~ factor(subreddit.name), data = posts_data)

Residuals:
    Min       1Q   Median       3Q      Max
-62.72  -25.42    0.00    0.00  2870.28

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.3333    69.5529   0.077   0.939
factor(subreddit.name)aaplwheel  -5.3333    184.0198  -0.029   0.977
factor(subreddit.name)alternativeeconomics -4.3333    184.0198  -0.024   0.981
factor(subreddit.name)amcstock    -4.3333    184.0198  -0.024   0.981

Residual standard error: 170.4 on 392 degrees of freedom
(9468 observations deleted due to missingness)
Multiple R-squared:  0.03266,    Adjusted R-squared:  -0.3103
F-statistic: 0.09523 on 139 and 392 DF,  p-value: 1
```

Data Visualization

Visualization Development

With our visualizations, we aimed to explore the timeline of events to provide context to the viewers, compare sentiment of posts and comments, analyze the difference in sentiment between subreddits, and discover the difference of words used between positive and negative sentiment comments.

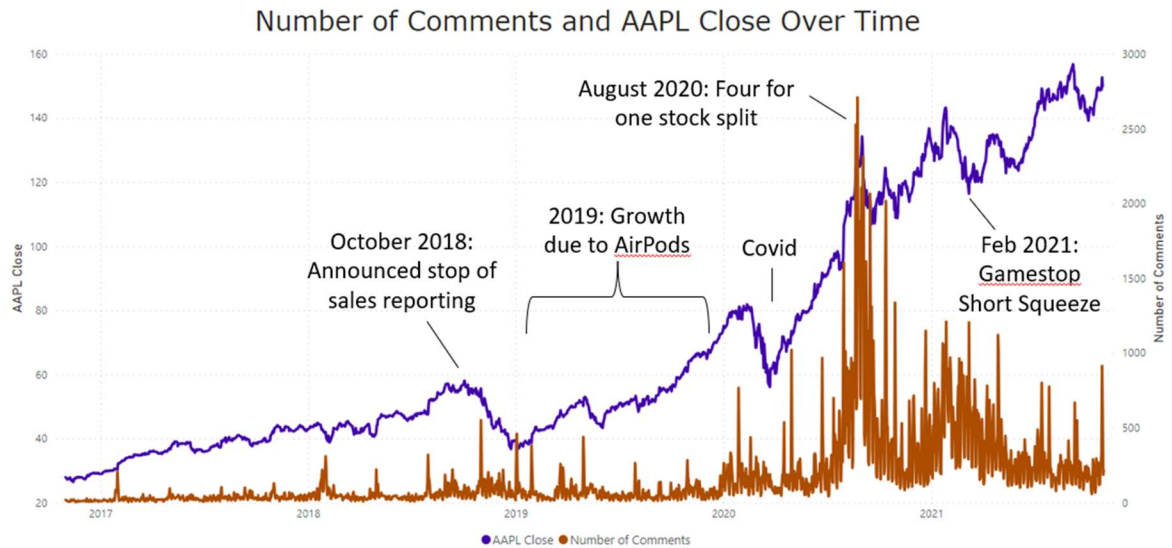


Figure 1: Number of Comments and AAPL Close Over Time

The first visualization used is a time series line graph, comparing the daily stock price of AAPL and the number of comments on Reddit that mentioned AAPL. This visualization was created in PowerBI and events were labeled later in PowerPoint. With this visual, we aimed to set the stage for the analysis later in the presentation. The different events on the timeline were explored, such as the dip in stock price in late 2018 and the huge spike in number of comments in mid 2020. Dark colors were chosen to view the details on the graph as light colors would have made this hard to view. The two lines also cross a few times, therefore, contrasting colors were chosen to be able to view the differences.

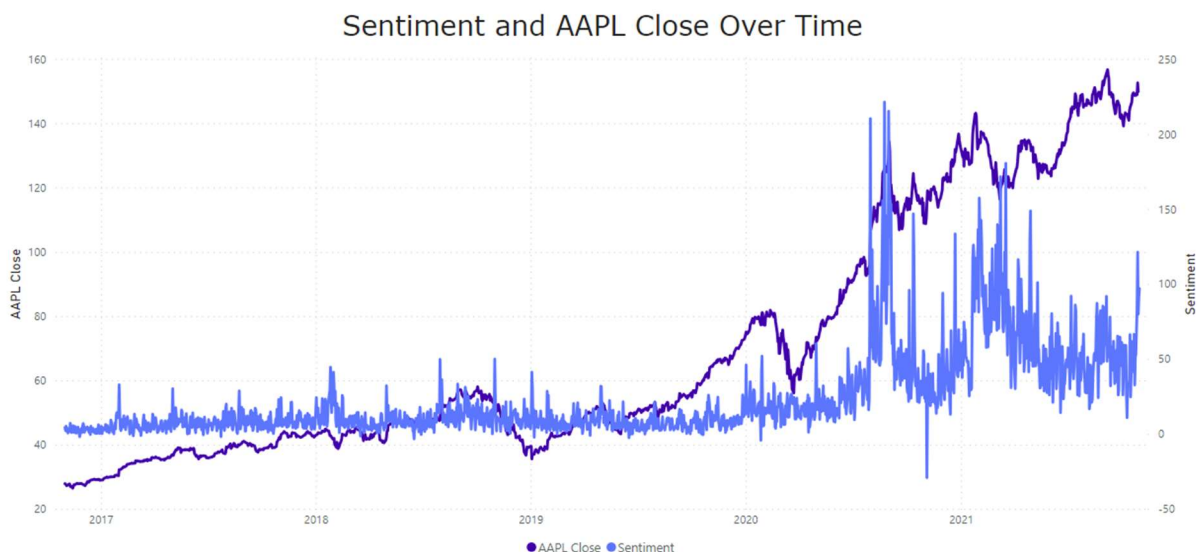


Figure 2: Sentiment of Comments and AAPL Close Over Time

The next visualization is also a time series line graph, displaying daily AAPL stock price and overall sentiment of Reddit comments that mention AAPL. This visualization was also made in PowerBI and is used to dive deeper into the sentiment of the comments and how they correspond with the events laid out in the previous visualization. Contrasting colors were chosen for the two lines because the lines overlapped a lot and contrasting colors make it much easier to distinguish between the two lines. A lighter color was chosen for sentiment but a deliberately medium tone was chosen. The color for the stock price was also kept the same from the previous graph to be consistent across slides.

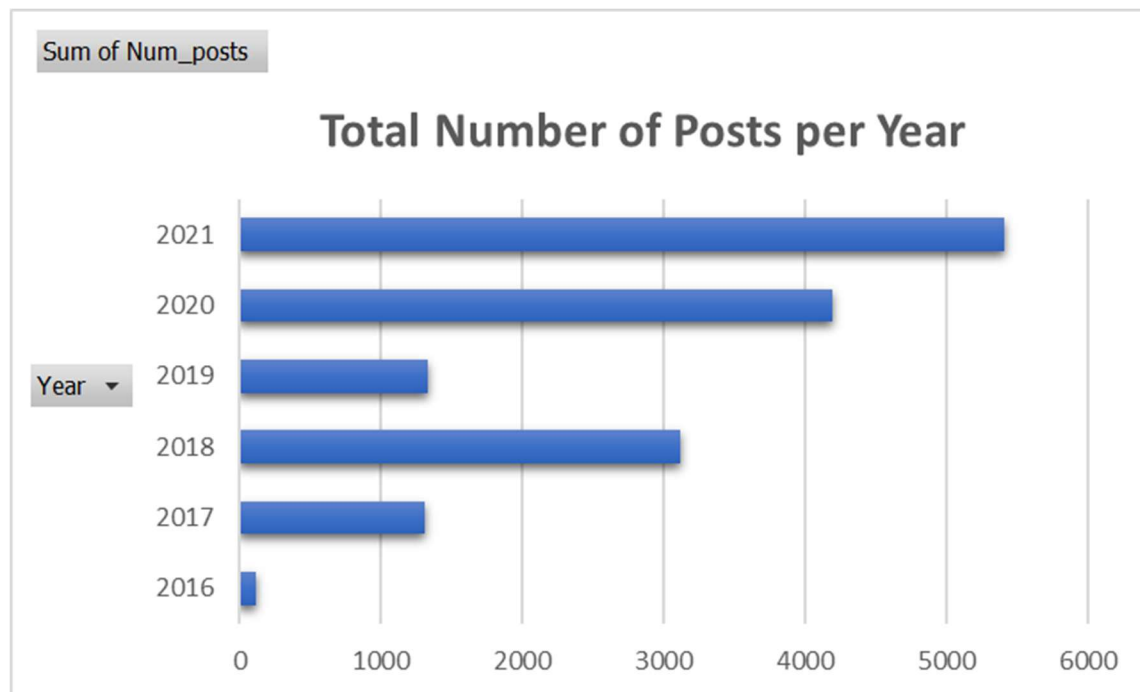


Figure 3: Number of Posts by Year

This visual displays the total number of Reddit posts per year to highlight the growing activity surrounding the stock market on just one social media platform. This bar chart visualization was created in R and focuses on all subreddits. One thing that should be noted is our dataset only captures up to October of 2021 so even without having 3 months of data, 2021 was the most active year in terms of posts mentioning Apple. This could have been portrayed using a line chart or even a pie chart, but the bold, wide shapes present the data in a more efficient and effective way.

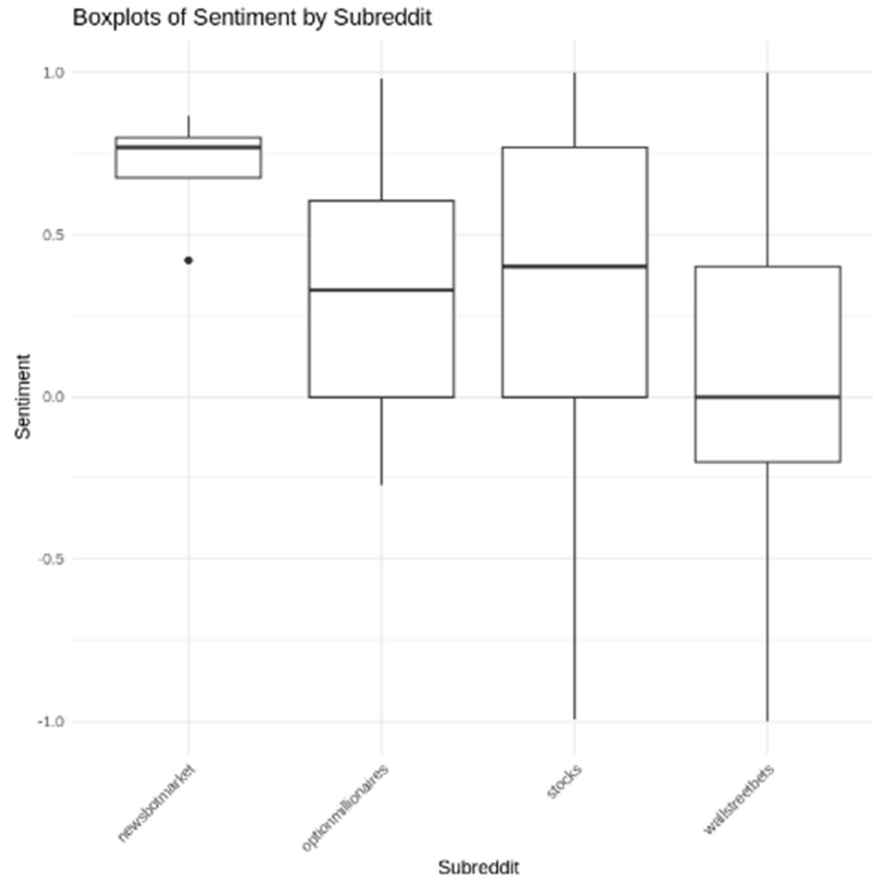
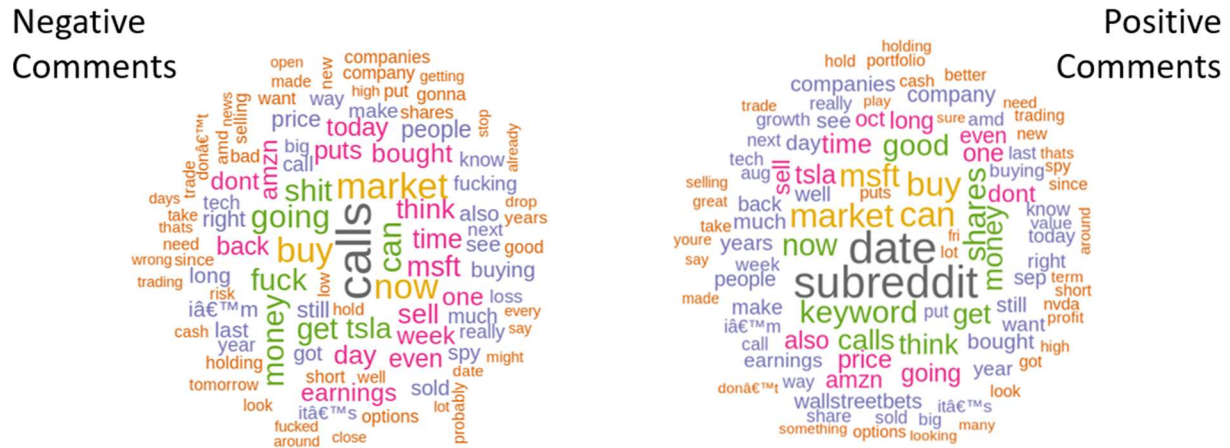
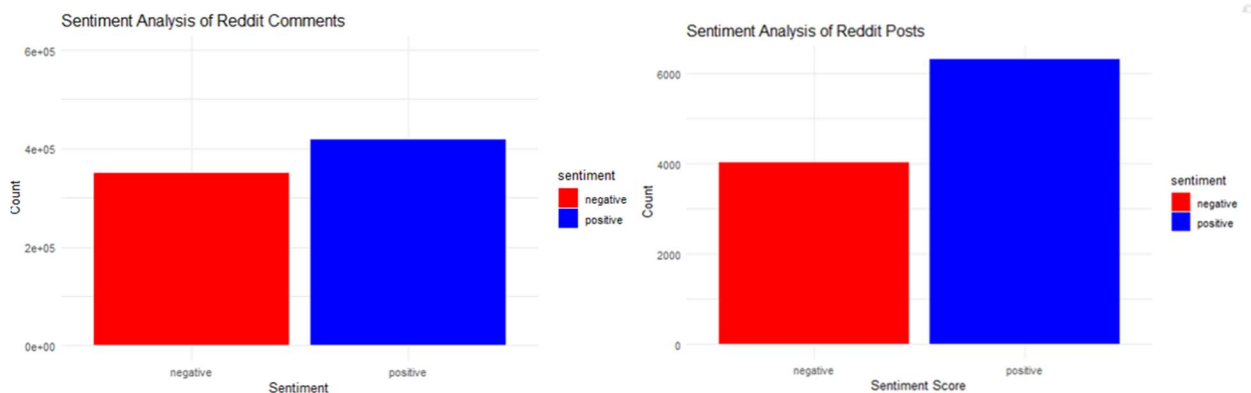


Figure 4: Spread of Sentiment by Subreddit

This visual displays the spread of comment sentiment for each of the top four subreddits where AAPL was mentioned the most. This boxplot visualization was created in R and displays how different types of subreddits can have very different sentiments towards the same topic. Even though there were hundreds of subreddits we could visualize, a subset of subreddits was displayed instead to make the chart easier to view and analyze. The y axis was set to only display values between -1 and 1 because the sentiment value for each comment was between these values.



These word clouds were created in R on Google Colab. The word cloud on the right displays the most frequently used words in the comments with a negative sentiment score. The word cloud on the left displays the most frequently used words in the comments with a positive sentiment score. Our text preprocessing process is described in the Text Analysis section of this paper. We followed each of these steps to ensure that the visualizations that we produced accurately represented our data.



Using R's packages, we were able to breakdown comments and posts into individual words to gauge the sentiment behind the content. With the `bing` sentiment lexicon, nearly 800,000 commented words were categorized as negative or positive. This provides a great understanding of the reputation surrounding Apple and how the largest company in the world is viewed by the public. The sentiment of comments is different than the sentiment variable in the comments dataset as the values were based on the entire comment rather than the individual word.

Visualization Interpretation

Figure 1: Number of Comments and AAPL Close Over Time

As we compare major events in Apple's stock history, we see that events in the news affect the price of the Apple stock. For example, after Apple announced they would no longer be releasing their quarterly sales in October 2018, the stock price decreased drastically. In parallel, we see the number of comments on Reddit about AAPL go up when a major event with Apple occurs. For example, when Apple announced they would be doing a four for one stock split, the number of comments on Reddit about AAPL skyrocketed. This chart shows that there is a relationship between AAPL stock price and the number of comments on Reddit about AAPL. This is the first step in determining whether Reddit data can be used to enrich stock analysis.

Figure 2: Sentiment of Comments and AAPL Close Over Time

Similarly to the previous time series chart, this graph shows that as events transpire and the AAPL stock changes drastically, the sentiment of comments also changes. For example, when the AAPL four for one stock split was announced, sentiment about AAPL in Reddit comments raised drastically. This chart further shows there seems to be a relationship between AAPL stock price and comments on Reddit.

Figure 3: Number of Posts by Year

The bar chart provides a quick and concise overview on the total number of posts per year that mention Apple. The main takeaway is the growing interest in the market by the average individual which has immensely increased since the pandemic and trends suggest that "retail investors", as they're referred to by financial market talking-heads, are only anticipated to grow. Whereas analysts can share their opinions on Bloomberg or on other forms of media, the average investor has a limited way to project their research and findings

Figure 4: Spread of Sentiment by Subreddit

This chart demonstrates the difference in spread of sentiment of comments in different subreddits. The purely news subreddit, r/newbotmarket has a much smaller spread with a higher mean, which means almost all posts in that subreddit are more positive than the other opinion based subreddits. r/optionmillionaires and r/stocks had a higher average sentiment than r/wallstreetbets, however, the spreads are almost as large. The mean sentiment of r/wallstreetbets is very neutral and the sentiment spreads from -1 to 1, the full range, which shows that this subreddit has the greatest variety in sentiment. Overall, these boxplots demonstrate that the subreddit where data is pulled can vary quite a bit in sentiment and this is something to be wary of when considering using this data to enhance stock analysis.

Figure 5: Word Clouds of Most Frequently Used Words in Positive and Negative Comments

These two word clouds show the differences between the frequently used words in positive comments compared to negative comments. Although the word clouds contain several similar words, as well as many words that do not immediately indicate positive or negative sentiment to the average person, there are some specific differences between the two. Specifically, the negative word cloud contained a few specific words that indicate loss, while the positive word cloud contained a few words that indicated growth. This makes sense when discussing stocks. Additionally, the negative word cloud contained several swear words. This comparison helps us to understand what types of words influence a comment's sentiment score, which then gives us a clearer understanding of why these scores may change in relation to other variables.

Figure 6: Sentiment of Comments and Posts

These two bar charts show the overall positive sentiment and negative sentiment of comments and posts. The positive sentiment of comments is only slightly larger than the negative sentiment. The positive sentiment of posts is much larger than the negative sentiment. This shows that posts are overall more positive than comments, which is important to consider if Reddit data is used to enhance stock analysis.

Conclusion

After completing our research project, did we find any unique relationships? For the most part, no. We found that the hypothesized correlation between social media sentiment and stock performance might not be as it seems. In fact, we found that stock performance is more likely to influence social media rather than the other way around. This is an important distinction to make as many algorithmic traders depend on sentiment data as a baseline for many of their trading bots. However, our research isn't perfect and there are many areas for future improvements.

First, we'd like to expand to analyze different sized companies. Does this same analysis work for a small company with a market capitalization below \$100M? \$10M? Does sentiment affect smaller companies more than larger companies? In addition, we'd like to analyze cross-platform data, and test hypotheses to see which platform sentiment data fits the stock data best. As a group we also discussed the potential to analyze trading volume data instead of stock price as it is potentially a closely related indicator. Finally, we'd like to cross analyze our data with news data and perform sentiment analysis on all stories on AAPL by day. It is well known that stock quarterly reports and new stories are heavily influential on investor reactions and trading, and we believe this would greatly enrich our dataset.

References

- “GameStop Short Squeeze.” *Wikipedia*, Wikimedia Foundation, 11 Dec. 2023, en.wikipedia.org/wiki/GameStop_short_squeeze.
- Gurman, Mark. “Apple Will Stop Reporting Unit Sales of iPhone, iPad and Mac from Next Quarter.” *HT Tech*, 2 Nov. 2018, tech.hindustantimes.com/tech/news/apple-will-stop-reporting-unit-sales-of-iphone-ipad-and-mac-from-next-quarter-story-eGHAZadU6zmvVCimfyRPKL.html.
- Kharpal, Arjun. “Apple Got off to a Strong Start and Looked Unstoppable in 2020. Then the Coronavirus Broke Out.” *CNBC*, CNBC, 10 Mar. 2020, www.cnbc.com/2020/03/10/coronavirus-apple-shares-looked-unstoppable-in-2020-until-the-outbreak.html.
- Lexyr. “Five Years of AAPL on Reddit.” *Kaggle*, 3 Nov. 2021, www.kaggle.com/datasets/pavellexyr/five-years-of-aapl-on-reddit.
- Noonan, Keith. “Why Apple Stock Soared 86.2% in 2019.” *The Motley Fool*, The Motley Fool, 12 Jan. 2020, www.fool.com/investing/2020/01/12/why-apple-stock-soared-862-in-2019.aspx.
- Volkman, Eric. “Apple Announces 4-for-1 Stock Split.” *The Motley Fool*, The Motley Fool, 31 July 2020, www.fool.com/investing/2020/07/30/apple-announces-4-for-1-stock-split.aspx.

Appendix

AAPL Stock and Reddit

Derek Bobbitt, Claire Plourde, Macy
Broderick, Matthew Clarke

\$GME: Social Media & Wall Street

Quick Class Poll: Who remembers the GameStop Saga?

- Brought social media to mainstream attention as a form of influence in financial markets. Culture Theory.
- In trading communities, it's now pretty common for algos and quants to include scraping SM sentiment in their trading bots
- Can we really say it is worth it to look at Reddit activity and sentiment when predicting stock performance?
- Goal: Explore the relationship of Reddit sentiment and stock performance, and vice versa.



\$GME: Social Media & Wall Street

< Class Poll: Who remembers the \$GME Saga? >

- ☐ The GameStop Saga was the first time social media was truly brought to mainstream attention as a form of influence in financial markets.
- ☐ Although it was common in day and quantitative trading communities to scrape social media data as part of their analysis or data for training trading bots, now analyzing social sentiment became a crucial step across most trading strategies even in corporate finance.
- ☐ The idea that a group can organize through social media, invest in sync or at the very least in tandem, had never been seen before at the scale seen with the r/wallstreetbets community, but now that is was a reality, we've seen this trend grow greatly.
- ☐ However, can we really say it is worth it to look at any one social media platform and find a statistically significant correlation when utilizing sentiment analysis?
- ☐ Our goal with this research project is to explore the relationship of Reddit sentiment and stock performance, and vice versa (important for later)

Our Data

- Datasets: 'Five Years of AAPL on Reddit' + 'AAPL Daily Stock Performance'
 - All AAPL mentions on Reddit, from Nov 2016 to Oct 2021
 - Data includes values that was useful for sentiment and time-based analysis
- Why enrich our data with AAPL stock performance?



Our Data

- ☐ Datasets: 'Five Years of AAPL on Reddit' + 'AAPL Daily Stock Performance'
 - All mentions of AAPL from Nov 2016 to Oct 2021
- ☐ Data includes values that was useful for sentiment and time-based analysis
- ☐ Why enrich our data with AAPL stock performance?
 - In addition to the Reddit dataset, we used a dataset labeled “AAPL Daily Stock Performance” from Yahoo Finance to enhance our existing data. The combination of the two datasets allowed us to analyze the relationship between Reddit posts and comments and Apple’s stock price over time.

Why these Datasets?

- Team interest from \$GME Saga and social sentiment effect on stock market
- AAPL is the largest company in the world ==> LOTS of coverage
- Would this apply to a popular comp?
- Dataset had necessary text and numeric data for our sentiment analysis – clear connection to AAPL stock data for enrichment

About this file
The table containing all the comments.

A type	id	subreddit.id	subreddit.name	subreddit.nsfw
Type of the datapoint	Unique Base-36 ID of the comment	Unique Base-36 ID of the comment's subreddit	Human-readable name of the comment's subreddit	Is the comment's subreddit NSFW
1 unique value	297533 total values	297533 total values	wallstreetbets stocks Other (96345)	61% 7% 32%
comment	hitr97r	2ujfk	stocks	false
comment	hitq83x	2ujfk	stocks	false
comment	hitp0ey	3q8bq	millennialbets	false
created_utc	permalink	body	sentiment	score
Timestamp of the comment's creation	Permalink to the comment on Reddit	Comment's body text	Analyzed sentiment for the comment	Comment's score
297533 total values	297533 unique values	288543 unique values		
1635724579	https://old.reddit.com/r/stocks/comments/qjv011/will_a_few_apl_shares_and_buy_come_hitr97r/	I own all 3. Don't sell AAPL.	0.0	1
1635724842	https://old.reddit.com/r/stocks/comments/qj87jj/on_texas_valuation/hitq83x/	I believe TSLA want to be like AAPL: part hardware and part software. I am unsure they will be and ...	-0.25	2
1635723463	https://old.reddit.com/r/MillennialBets/comments/qk1p67/charlie_munger_doubled_own_on_babe_for_a_co...	***[Recent News for BABE-J] (https://www.reddit.com/r/MillennialBets/wiki/index/stocks/BABA)** Date: 11...	0.9422	1

Why These Datasets?

- ☐ Team interest from \$GME Saga and social sentiment effect on stock market
- ☐ As for the scope of our research, we chose to investigate if sentiment could really move some of the world's largest companies or if GME was the exception, not the rule. Accordingly, we chose to look at Apple, the largest company in the world. This means that there's a lot of coverage and an abundance of data.
- ☐ AAPL is the largest company in the world ==> LOTS of coverage
- ☐ We asked: would this kind of social sentiment pressure apply to a popular comp?
- ☐ Dataset had necessary text and numeric data for our sentiment analysis – clear connection to AAPL stock data for enrichment

Data Journey *Contents and Management*

Dataset contained two csv files:

- Comments.csv
 - 297,533 rows
- Posts.csv
 - 15,483 rows

Comments Data Dictionary	
Name	Description
type	Type of the datapoint
id	Unique Base-36 ID of the comment
subreddit.id	Unique Base-36 ID of the comment's subreddit
subreddit.name	Human-readable name of the comment's subreddit
subreddit.nsfw	Is the comment's subreddit NSFW?
created_utc	Timestamp of the comment's creation
permalink	Permalink to the comment on Reddit
body	Comment's body text
sentiment	Analyzed sentiment for the comment on score from -1 to 1
score	Comment's score

The dataset we chose consisted of 2 files. One contained data on the Reddit comments mentioning the AAPL stock ticker. This file had 297,533 rows of data. The other file contained data on the Reddit posts with mentions of the AAPL stock ticker. This file had 15,483 rows of data.

The information available to us in these datasets included things like the subreddit name, the timestamp, the body text of the comments and posts, etc. The data dictionary for the comments file can be seen on this slide. The posts dataset followed mostly the same structure with similar information available.

Data Journey

Preparing the Data

Data Cleaning:

- Created_utc
- NSFW
- Subreddit.name == '???'
- Sentiment & Score

```
# Check for duplicate rows
duplicates_comments = comments %>%
  distinct() %>%
  filter(duplicated(comments))

# Print the number of duplicate rows
print(nrow(duplicates_comments))
```

```
#Convert all text to lowercase for analysis
posts = posts %>%
  mutate(selftext = tolower(selftext)) %>%
  mutate(title = tolower(title))

comments = comments %>%
  mutate(body = tolower(body))
```

```
#add length of comment
comments = comments %>%
  mutate(comment_length = sapply(strsplit(body, " "), length))
```


Our dataset came very clean, but in order to clean it further we began in excel with the following steps:

- created_utc (timestamp) was originally a number, so we had to use a formula to change this to date time format.
- NSFW (if the subreddit is qualified as not safe for work) was originally true or false. We used find and replace to change these values to 1 and 0.
- Subreddit.name listed as '???' we looked up this line of data based on the other categories such as the link to find the actual name of the subreddit and replaced it.
- Sentiment and Score columns had to be changed to numeric data types.

Further, in R we checked for duplicate rows, converted all text data to lowercase, and added a column calculating the column length which would be used in our later analysis.



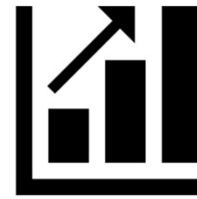
Data Journey *Challenges and Triumphs*

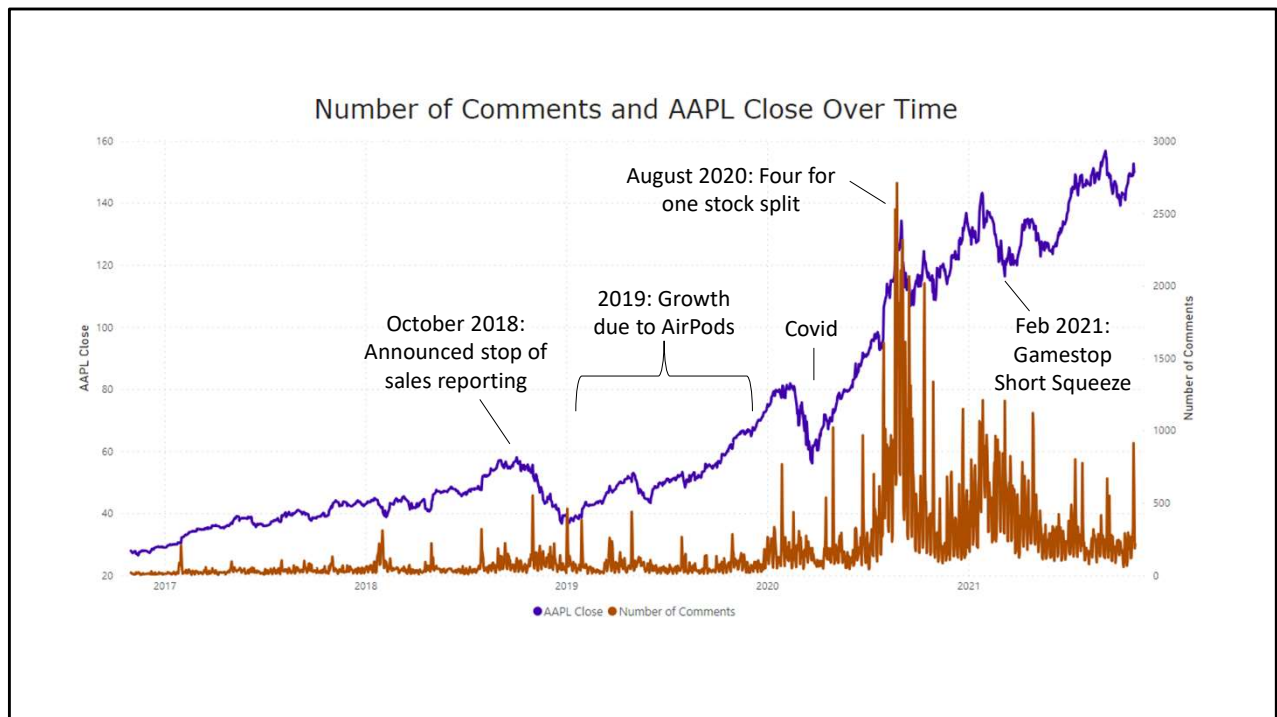
- Missing Values
 - Enhancing our Data
 - Making Data Uniform
 - Datatypes
- 

Some of the challenges we ran into in this process included:

- Missing values - Specifically in the sentiment score column, there were quite a few rows with missing data. We added in 0 for these columns, a neutral value, as not to affect the analysis.
- Enhancing our data - The information available in our dataset was slightly limited due to the low number of usable columns in each dataset. Because of this, we decided to enhance our data with additional information about current events at the time of certain trends in our data.
- Making data uniform for usability - This wasn't too challenging, but it did require a little bit of additional work to make sure all of our data was in the proper format and prepared for analysis so it would be accurate.
- Datatypes - We had to change some of the datatypes, such as the timestamp, to make sense of them in terms of the analysis. Additionally, some of the datatypes, such as the ones with character values, were more difficult to use in our analysis in terms of quantifying.

Exploratory Data Analysis

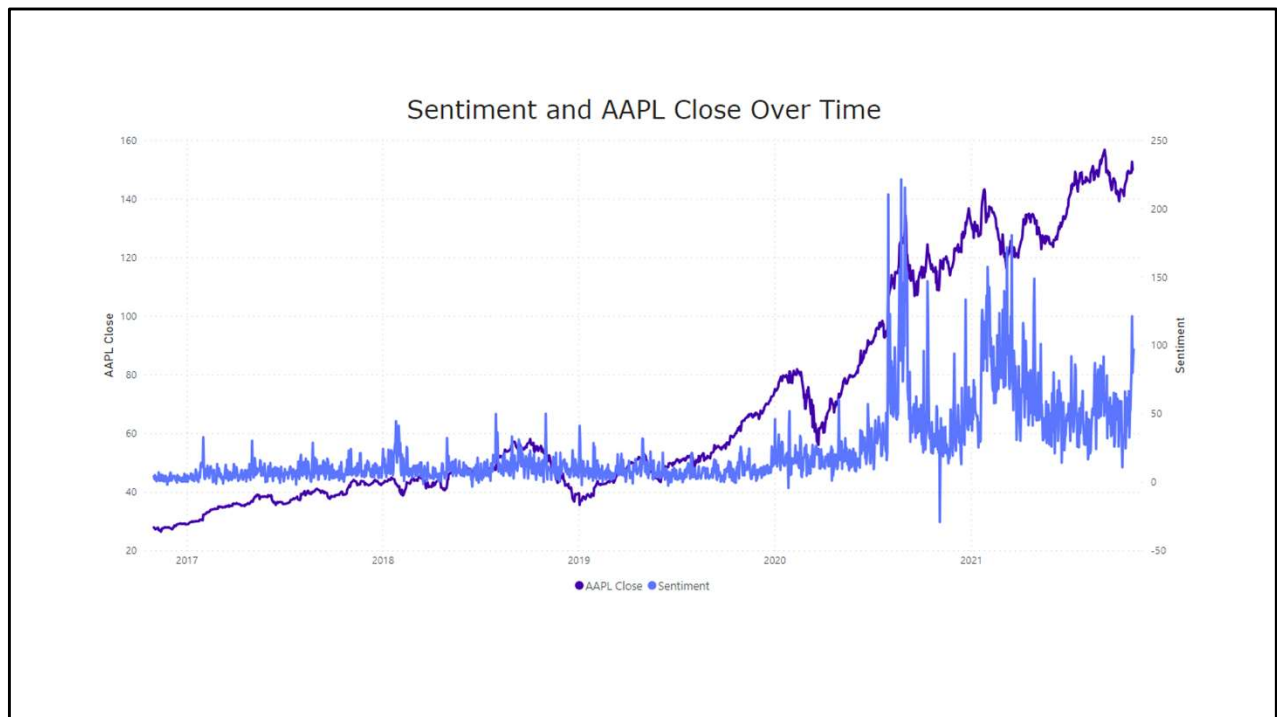




- Purpose: set up a timeline of events
- Takeaway: ups and downs in AAPL stock, Reddit comments respond to events
- AAPL stock price and number of comments about the AAPL stock over time
- Looking at the trend of the stock, there are a few key events
- In October 2018, Apple announced they would stop reporting sales of iPhone, iPad, and Mac
 - Seen by the public as a way to hide any struggles they were having, stock took a nosedive, lots of chatter on Reddit
- In 2019, the second gen of AirPods was a huge success, which translated into growth for the company
- Early 2020 drop due to Covid
- At the end of August 2020, Apple announced a four for one stock split (investors will get three shares for every share they hold)
 - Caused a lot of people to buy and then sell shortly after
 - Also explains a lot of activity on Reddit that day, people providing advice, telling other people to buy
- Late January 2021: Gamestop Short Squeeze
 - Short sellers borrow shares and immediately sell them, hoping to buy them back at a lower price later, return them to the lender at a profit
 - This practice is common on stocks that don't have a lot of value, one of which is

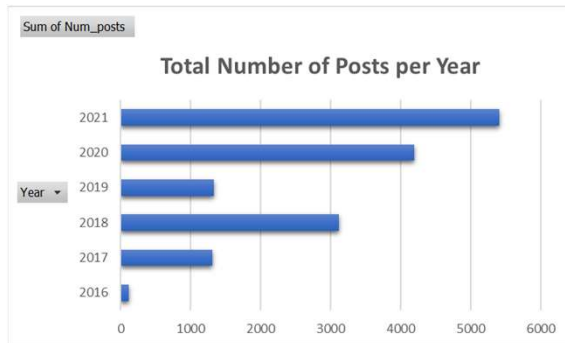
Gamestop, a once popular but now dying video game company

- The Reddit community wallstreetbets decided to short squeeze this stock and other low value stocks by driving up the price and causing huge losses for short sellers
- AAPL was not one of these stocks but this movement caused a small economic panic in the stock market
- We can see here that this panic did affect the AAPL price
- The Gamestop short squeeze had lasting effects on Reddit chatter around stocks: more comments than ever before



- Sentiment vs AAPL Stock price over time
- Sentiment was pretty neutral, slightly positive until August 2020
 - It is clear people really loved Apple's idea to do a four for one stock split, sentiment was hugely positive
- Goes into a huge dip in late 2020 (November 4th)
 - Apple had dropped for the second time in a few weeks
 - The election of 2020 (Biden v Trump) happened on November 3rd, a lot of tension
- Very positive from that moment on
 - Another spike in March 2021, had been falling for a while, a lot of people happy about the lower price
- Even though APPL was not directly affected by the Gamestop Short Squeeze, it got a lot more people interested in investing, which led to more conversation and an overall positive sentiment
- Macy and Matt will go into specifics later about what these posts are talking about and focusing on

Summary Statistical Analysis



Posts per Subreddit

wallstreetbets	3730
stocks	865
optionmillionaires	710
newsbotmarket	675
forexhome	560
options	556
newsbotbot	462
investing	438
ultraalgo	407
xtrades	235

Average Comment Sentiment:

0.155344

Average Comment Score:

5.19916

Average Post Score:

27.3624

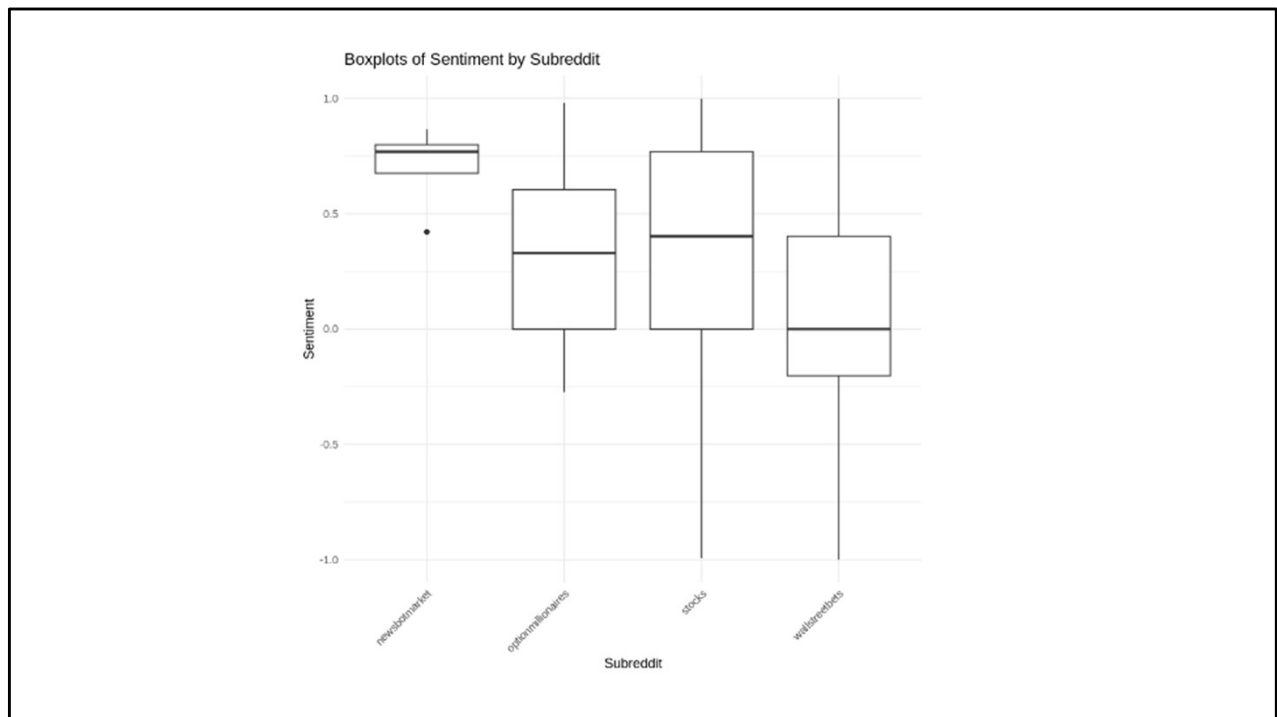
Purpose: Provide an overview of the dataset through summary statistics

Takeaway: Reddit's popularity

On the right side, the mean of the comment sentiment, comment score, and post score are displayed to quickly inform the viewer of the important variables that this project focuses on.

- The score variable is subjective in the sense that users decide whether they like or dislike though controls called "upvotes" and "downvotes"
- The data suggests that the average comment's sentiment is slightly positive, which for a public forum is good to note. Some consideration should be made since the assumptions from the word dictionary cannot be observed.
- Next to the average are two tables overlayed highlighting the total posts per subreddit and comments per subreddit
- These data points highlight the potential for cultural theory to be observed as each subreddit has a different culture and having the majority of data points from one subreddit could influence the results.
- Wallstreetbets was the most active subreddit in terms of posts and comments about Apple, surprising considering Apple is assumed to be a safe stock, and Wallstreetbets is known for more reckless investing.
- And finally, a bar graph highlighting the total number of posts per year.

- The most interesting part of the graph comes from the 2021 column as the dataset only recorded up to October 2021 so in 9 months, Apple was mentioned in more posts compared to other years.
- This data provides a concise overview on what we have explored through data analysis



Purpose: break down sentiment by subreddit

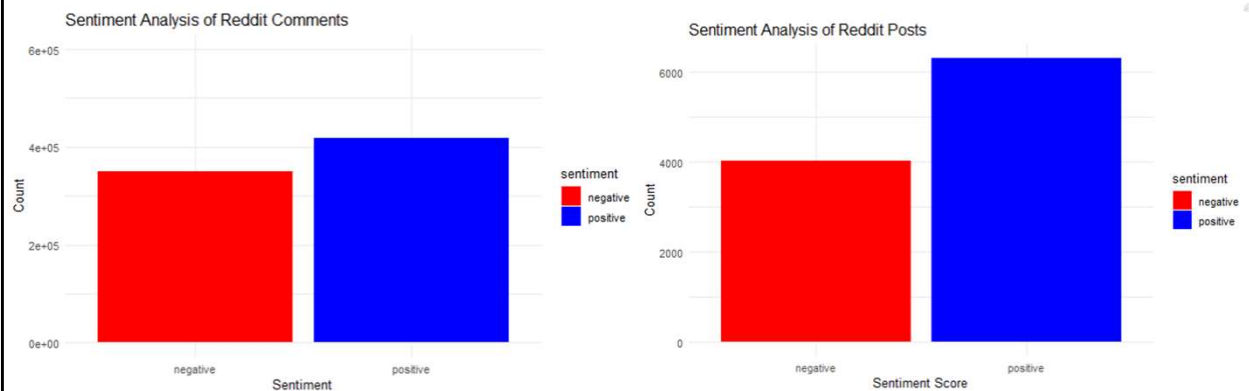
Takeaway: Different types of subreddits present different opinions on a similar subject and where you choose to get your information will affect the sentiment of the information

Next, we will look at sentiment broken down by subreddit

- Top 4 subreddits based on activity
- Newsbotmarket: news based subreddit, not a lot of opinion
- Optionmillionaires, stocks, and wallstreetbets: dedicated to stock discussion, what stocks to invest in, recent trends, a lot more opinion based
- The news based subreddit has the highest average sentiment, does not have a large spread
- The opinion based subreddits have a lower average sentiment and a larger spread
- Wallstreetbets has a lower, neutral average, probably due to lots of positive sentiment and lots of negative sentiment
- Stocks and optionmillionaires have a higher average than wallstreetbets
- If you're looking for strictly positive outlook on stocks, you should browse newbotmarket
- If you're looking for a large spread in opinions but still overall positive, view stocks or optionmillionaires
- If you're looking for the most extreme opinions, browse wallstreetbets

- Where you get your information matters!

Sentiment – Posts v Comments



Purpose: To show the difference in sentiment between comments and posts

Takeaway: There were more positive than negative sentiment in each group, and comments clearly had the larger dataset, but posts had a better positive-negative ratio than in comments

Next, we will look at sentiment broken down by positive and negative

- There were over 400,000 positive words in the comments dataset, whereas posts contained just above 6,000
- The Positive-negative ratio for posts were roughly 3:2, whereas for comments it was ~4:3
- A possible explanation for the higher ratio is due to the higher standards held for posts; nearly all subreddits have one or many moderators to review the content people want to share in the community, and that process is not applied to comments
- So comments can naturally be more raunchy, hateful, and longer than posts
- Additionally, the general rule for investing is to buy a stock and hold in hopes of the price to increase so it is unlikely that people will post negative things regarding Apple if they are also invested in the company, and the average retail investor is not exposed to shorts.

Multiple Linear Regression

- To see the influence of other variables on the sentiment score of comments and posts

```
Call:
lm(formula = sentiment ~ comment_length + factor(subreddit.name),
    data = comments_data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.32716 -0.26746 -0.04279  0.36418  1.17272

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.855e-01  3.253e-01   1.800  0.071940 .
comment_length 1.037e-03  6.888e-05  15.061 < 2e-16 ***
factor(subreddit.name)4chan  8.444e-02  5.634e-01   0.150  0.880865
factor(subreddit.name)4kto1m  2.825e-01  5.634e-01   0.501  0.616074
factor(subreddit.name)5kto100kchallenge -2.845e-01  5.634e-01  -0.505  0.613641
factor(subreddit.name)aapl -3.771e-01  3.429e-01  -1.100  0.271478
factor(subreddit.name)aaplwheel  2.192e-01  5.634e-01   0.389  0.697285
factor(subreddit.name)accounting -6.706e-01  5.634e-01  -1.190  0.234006
factor(subreddit.name)activeoptiontraders -3.490e-01  5.634e-01  -0.619  0.535675
factor(subreddit.name)algotrading -2.839e-01  3.494e-01  -0.812  0.416552
factor(subreddit.name)alpp -1.721e-01  5.634e-01  -0.305  0.760067
factor(subreddit.name)ama  8.821e-02  4.600e-01   0.192  0.847948
factor(subreddit.name)amcstock -4.320e-01  3.514e-01  -1.230  0.218858

Residual standard error: 0.46 on 8621 degrees of freedom
(942 observations deleted due to missingness)
Multiple R-squared:  0.1533, Adjusted R-squared:  0.1105
F-statistic: 3.58 on 436 and 8621 DF, p-value: < 2.2e-16
```

```
Call:
lm(formula = score ~ factor(subreddit.name), data = posts_data)

Residuals:
    Min       1Q   Median       3Q      Max
-62.72  -25.42    0.00    0.00  2870.28

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3333    69.5529   0.077  0.939
factor(subreddit.name)aaplwheel -5.3333    184.0198  -0.029  0.977
factor(subreddit.name)alternativeeconomics -4.3333    184.0198  -0.024  0.981
factor(subreddit.name)amcstock -4.3333    184.0198  -0.024  0.981

Residual standard error: 170.4 on 392 degrees of freedom
(9468 observations deleted due to missingness)
Multiple R-squared:  0.03266, Adjusted R-squared:  -0.3103
F-statistic: 0.09523 on 139 and 392 DF, p-value: 1
```

Purpose: See if there was significant influence from a variable on sentiment and score

Takeaway: Only comment length and a few subreddits showed statistical significance when using the comments data, and nothing was reported when looking at score in the posts data

- R^2 measures the percent of variability in Y that is explained in the regression
- Both had really low R^2 values, 0.1533 and 0.03266 respectively, and is not considered accurate enough to draw conclusions from.
- The more popular/active subreddits had slight significance for comments, but would be validated with a higher R^2
- Very cool to draw from the other MSBA class to see the applicability of these models
- Very high SE for factor(Subreddit name) in posts multiple linear regression model

Conclusion + Improvements

- Unique relationships? For the most part, no.
 - Might be that stock performance influences social media rather than the other way around for a company
- Future Improvements:
 - Analyze different sized companies - does this same analysis work for a small company with a Mkt Cap below \$100M? \$10M? Does sentiment affect smaller companies more than larger companies?
 - Cross-Platform Data – test hypotheses to see which platform sentiment fit the stock data best?
 - Instead of Stock Price, look at Volume as a closely-related Indicator?
 - Cross-Analyze with News Data – analyze text for all stories on AAPL by day



Questions?

Thank you!