BANA 277 Customer Social Analytics Project Paper: Wall Street Bets Reddit Community Sentiment & Social Analysis

Team 10: Vincent Lee, Bangyang Qiu, Wenchi Tseng, TJ Tolliver, Veronica Tseng

March 19th, 2024

Table of Contents

Project Introduction

In the digital age, the influence of social media on various sectors has been profound, with the financial markets being no exception. A notable example of this influence is the Wall Street Bets (WSB) community on Reddit, a forum that has gained notoriety for its significant impact on stock prices through collective investment strategies. This project centers on analyzing the effects of the Wall Street Bets Reddit community on the stock market. By leveraging a comprehensive dataset obtained from Kaggle, which encompasses posts from the WSB community, we aim to uncover how sentiments and discussions within this vibrant forum correlate with movements in stock prices.

Our project examines the Wall Street Bets Reddit community's influence on stock markets, focusing on the GameStop phenomenon in 2021. Using a dataset with variables like post titles, upvote scores, and comment counts, we apply data visualization, sentiment, correlation, and regression analyses to explore how WSB discussions affect stock prices. This study highlights the power of retail investor sentiment, driven by social media, in influencing market dynamics and investment strategies. By analyzing WSB's impact, we aim to understand social media's role in shaping financial trends and the implications for both retail and institutional investors.

Wall Street Bets Community Context

The Wall Street Bets (WSB) community, a vibrant and dynamic forum on Reddit, represents a paradigm shift in how retail investors interact with the stock market. Originating as a space for sharing investment tips, strategies, and discussions, WSB has evolved into a significant force capable of mobilizing collective actions that impact financial markets. Known for its candid and often humorous approach to trading, the community gained international fame in January 2021 with the GameStop saga, where its members allegedly drove up the stock's price from around $4 at the beginning of the month to a peak of $80 at the end of the month. This marked a monumental event in financial history, challenging traditional market dynamics and hedge fund strategies. In fact, the price surge was so unprecedented that one of the forum members in particular faced a lawsuit for market manipulation as his posts received the most attention during the time.

By exploring the characteristics that make WSB unique, including its language, memes, and investment philosophies, we delve into how a social media forum has become a noteworthy player in the financial world. The GameStop event not only underscored the community's ability to affect stock prices but also raised questions about the role of social media in finance, market manipulation, and the democratization of investing. As we navigate through the WSB community context, we will uncover the

implications of its actions on market regulations, investor behavior, and the future of trading.

Exploratory Data Analysis

The exploratory data analysis reveals significant insights into the nature of the dataset obtained from Kaggle, focusing on the Wall Street Bets (WSB) Reddit community and its influence on stock markets, particularly during the GameStop phenomenon of 2021. With over 34,000 rows of posts and 8 variables, the dataset from Kaggle showcases a substantial volume of user-generated content. Notably, a considerable portion of posts lacks body content, likely due to the prevalent practice among Reddit users of predominantly sharing titles without additional context. This observation underscores the brevity and succinctness of communication within the WSB community, with an average post title length of approximately 10 words. In contrast, posts with body content exhibit a more typical average length of around 100 words, indicating a greater depth of discussion within those threads. Furthermore, the analysis highlights the presence of outliers in the upvote count variable, which substantially inflates the average to around 1,300, suggesting the existence of viral posts. However, it's crucial to recognize that such outliers skew the perception of typical post popularity, implying that the actual average upvote count per post is likely lower than initially indicated.

Sentiment Analysis

The method of analysis we chose to conduct for our project primarily involves sentiment analysis of the dataset. This aids in addressing one of our core questions in assessing the correlation between the community and stock prices. For example, by using sentiment classification, we can attempt to address whether or not there is sufficient evidence to suggest that one of the community members caused market manipulation of the GME stock price as mentioned earlier.

To apply sentiment analysis to our dataset, we first combined our two main text columns, "title" and "body", into a new column called "overall text." This allowed us to differentiate each data row under one column as an entire post's content instead of having to apply classification on two different columns. In our next step, we chose to utilize the VADER package through Python to classify each post as a sentiment between positive, neutral, and negative. Shown below is the preview of our transformed dataset that is ready to be used for sentiment analysis.
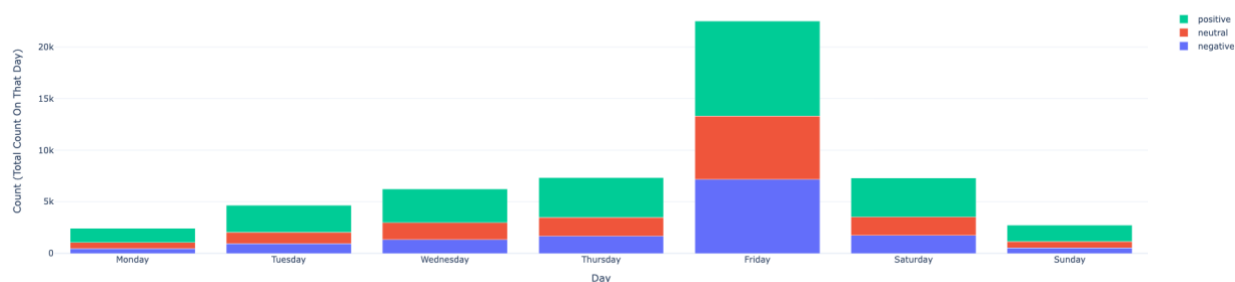
overall text	sentiment_scores	sentiment_category
It's not about the money, it's about sending a...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	neutral
Math Professor Scott Steiner says the numbers ...	{'neg': 0.287, 'neu': 0.713, 'pos': 0.0, 'comp...	negative
Exit the system The CEO of NASDAQ pushed to ha...	{'neg': 0.071, 'neu': 0.88, 'pos': 0.049, 'com...	negative
NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...	{'neg': 0.256, 'neu': 0.594, 'pos': 0.15, 'com...	negative
Not to distract from GME, just thought our AMC...	{'neg': 0.0, 'neu': 0.888, 'pos': 0.112, 'comp...	positive

Sentiment Visualizations

After applying sentiment classification on every row in our dataset, we were able to generate new data frames. Shown below are word cloud visualizations representing the dataset containing only positive sentiments and another dataset containing only negative sentiments. It is interesting to note that our classification picked the term "hedge fund" as associated with the negative sentiment word cloud. This ties in closely with the fact that the word "GME" is one of the popular words in both word clouds. During the GME price surge, hedge funds were considered an antagonist of the story as the community essentially drove the price of GME in retaliation against the hedge fund's short positions.

Word Cloud of Positive Sentiment Dataset

Word Cloud of Negative Sentiment Dataset

Another visualization shown below that we implemented after sentiment classification was a bar graph showing the distribution of posts and sentiments across the weekday. We can conclude that the total number of posts and sentiments is the highest on Friday. This is reasonable as Friday is the day when options and futures contracts expire, which may incite a lot of discussion posts that day. However, regardless of what day of the week it is, the distribution of sentiments is fairly evenly distributed. For example, on each day of the week, there is an even proportion of negative, neutral, and positive sentiments, so there is no evidence to suggest that a particular day will incite a particular sentiment.

WSB vs GME Analysis

Going back to the question of how the WSB community allegedly influenced the stock price surge of GME and whether a single forum member caused the entire movement, the question can be addressed with statistical insights from our sentiment analysis. Shown below is a pair plot with the top graph representing the price chart of GME downloaded using the yfinance package in Python and the bottom representing a

graph of the community sentiment. The sentiment graph was created by extracting all posts in the forum that contained the forum member's username. As a result, we are able to visualize the sentiment of the user's posts compared with the stock price of GME during the period of its price surge.

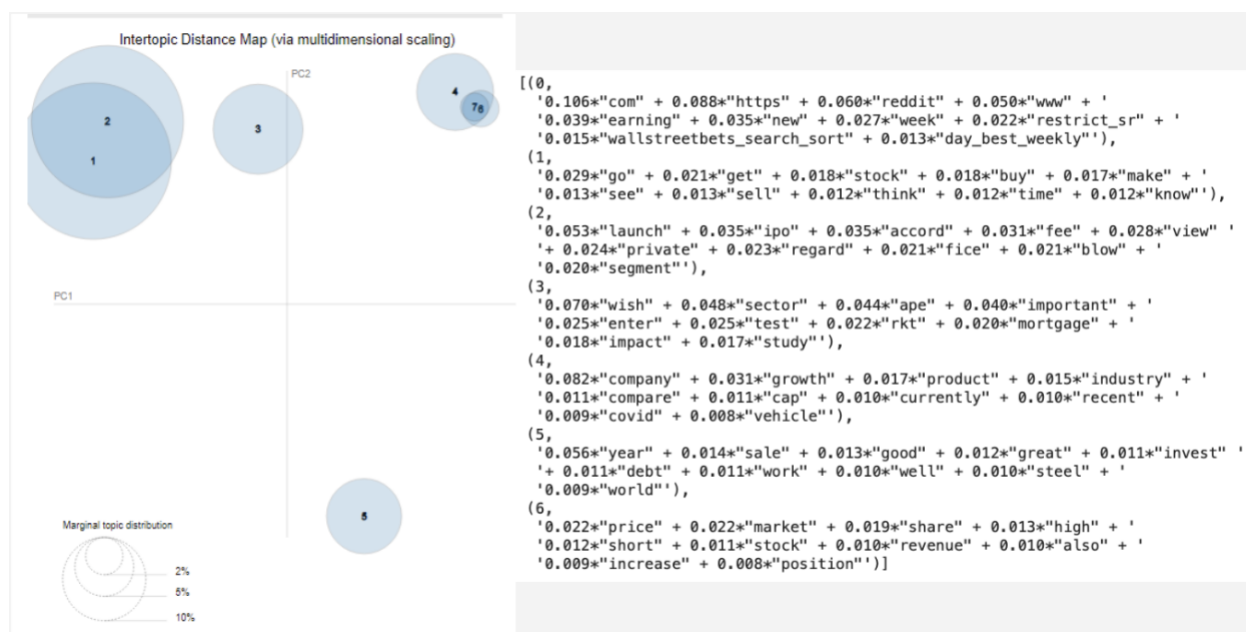GME Price and Majority Sentiment (Jan 2021 - April 2021)

At first glance, there is indeed a visual correlation between community sentiment and the initial price spike of GME. However, as we zoom out on the timeframe, the sentiment graph does not effectively capture the stock price movements in the later periods. In our statistical analysis, we applied correlation and regression analyses using the sci-kit-learn package in Python to compare the two graphs. The results are summarized below.

Correlation Coefficient	0.1093
Linear Regression MSE	368.6815
Linear Regression R-squared	-0.0572

Based on our results, we conclude that there is no statistical significance to suggest that the sentiment of the WSB community alone was influential enough to cause the stock price surge of GME. In an attempt to reject this conclusion, we applied other rigorous regression methods, such as the XGBoost Regressor, but they also produced similar results showing no statistically significant relationship between the sentiment and stock price.

Topic Modeling

On top of our sentiment analysis, we used the transformed dataset and applied topic modeling to better understand the dataset. Topic modeling is an unsupervised machine learning algorithm that clusters the data into topics based on word frequency and weights. To implement topic modeling, we converted our "overall text" into a list of individual words. Then we used lemmatization which converts every word to its root form for better implementation of the model. Finally, we chose to create 7 clusters in our topic modeling output. As a result, we were able to extract distinct topics ranging from trading information, stock news, company earnings, and so on. This allowed for a better understanding of the topics that the WSB community encompasses.


```
[(0,
 '0.106*"com" + 0.088*"https" + 0.060*"reddit" + 0.050*"www" + '
 '0.039*"earning" + 0.035*"new" + 0.027*"week" + 0.022*"restrict_sr" + '
 '0.015*"wallstreetbets_search_sort" + 0.013*"day_best_weekly"'),
(1,
 '0.029*"go" + 0.021*"get" + 0.018*"stock" + 0.018*"buy" + 0.017*"make" + '
 '0.013*"see" + 0.013*"sell" + 0.012*"think" + 0.012*"time" + 0.012*"know"'),
(2,
 '0.053*"launch" + 0.035*"ipo" + 0.035*"accord" + 0.031*"fee" + 0.028*"view" '
 '+ 0.024*"private" + 0.023*"regard" + 0.021*"fice" + 0.021*"blow" + '
 '0.020*"segment"'),
(3,
 '0.070*"wish" + 0.048*"sector" + 0.044*"ape" + 0.040*"important" + '
 '0.025*"enter" + 0.025*"test" + 0.022*"rkt" + 0.020*"mortgage" + '
 '0.018*"impact" + 0.017*"study"'),
(4,
 '0.082*"company" + 0.031*"growth" + 0.017*"product" + 0.015*"industry" + '
 '0.011*"compare" + 0.011*"cap" + 0.010*"currently" + 0.010*"recent" + '
 '0.009*"covid" + 0.008*"vehicle"'),
(5,
 '0.056*"year" + 0.014*"sale" + 0.013*"good" + 0.012*"great" + 0.011*"invest" '
 '+ 0.011*"debt" + 0.011*"work" + 0.010*"well" + 0.010*"steel" + '
 '0.009*"world"'),
(6,
 '0.022*"price" + 0.022*"market" + 0.019*"share" + 0.013*"high" + '
 '0.012*"short" + 0.011*"stock" + 0.010*"revenue" + 0.010*"also" + '
 '0.009*"increase" + 0.008*"position"')]
```

Conclusion/Further Research

In conclusion, our analysis of the Wall Street Bets subreddit sheds light on several key findings regarding its sentiments, correlation with stock prices, and community dynamics. We observe a relatively even distribution of sentiments across different days, with a notable concentration on Fridays, suggesting potential patterns in user engagement or market activity. However, despite apparent correlations between subreddit discussions and stock price movements, our investigation does not yield

conclusive evidence of a strong causal relationship. There are definitely arguments to be made regarding the fact that the WSB community influenced many retail traders to buy into the GME stock during that price spike period, and that in itself is a form of social influence. However, measuring that influence is very difficult especially when the financial market is so vast compared to the single social media forum community of WSB.

Further studies could utilize different forms of sentiment analysis and classification, and apply different machine learning or statistical techniques to extract causation and correlation insights. Additionally, our exploration reveals the diverse nature of the community, as evidenced by distinct topics emerging through topic modeling, ranging from trading advice to company analysis and news discussions. This diversity underscores the multifaceted nature of discussions within the subreddit, reflecting the varied interests and expertise of its members. Moving forward, further research and analysis are warranted to deepen our understanding of the interplay between online communities like Wall Street Bets and financial markets, providing valuable insights for investors and researchers alike.

References

Preda, G. (2021, August 16). *Reddit wallstreetbets posts*. Kaggle. https://www.kaggle.com/datasets/gpreda/reddit-wallstreetsbets-posts

wallstreetbets. Reddit. (n.d.). https://www.reddit.com/r/wallstreetbets/

Warren, K. (2023, October 13). *Who is Keith Gill? the man behind roaring kitty*. Investopedia. https://www.investopedia.com/keith-gill-roaring-kitty-8303143