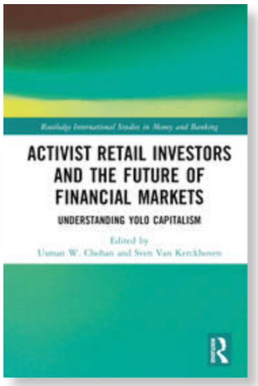
COSC2671: ANALYZING AND TRACKING THE SENTIMENT OF INVESTMENT SUBREDDITS FROM JULY TO AUGUST OF 2023

By Amay Viswanathan Iyer - s3970066

August 31, 2023

1: Introduction

The financial arena is experiencing tectonic shifts, accentuated by the dislocations caused by the COVID-19 pandemic. (1) The democratization of financial dialogue through digital platforms has markedly accelerated this transformation, questioning the historical dominion of institutional finance. (1) This newfound influence reached a watershed moment with the r/wallstreetbets subreddit's orchestration of a short squeeze on GameStop (GME), leading to significant financial setbacks for institutional giants like Melvin Capital. The forum's guiding principle, described as "Like 4chan found a Bloomberg Terminal," encapsulates the potent mix of financial savvy and social dynamics present in these online spaces. (1)


Routledge International Studies in Money and Banking
ACTIVIST RETAIL INVESTORS AND THE FUTURE OF FINANCIAL MARKETS
UNDERSTANDING YOLO CAPITALISM
Edited by
Usman W. Chohan and Sven Van Kerckhoven
R

To probe this emergent landscape empirically, this academic endeavor employs natural language processing (NLP) techniques to scrutinize both sentiment and prevailing conversational themes across 11 essential financial subreddits. (2) These forums represent a broad array of investment philosophies, from the traditional risk-averse stances found in 'r/investing' to the more speculative outlooks featured in 'r/wallstreetbets.' The study aims to offer an extensive and nuanced financial sentiment barometer based on 1,200 posts collected during a specific period, from the end of July to the end of August 2023.

It is crucial to emphasize that the findings and interpretations contained in this report should not be construed as financial advice. The analysis is context-specific and is confined to the sentiments expressed within the corpus of 1,200 posts studied.

1.1: Problem Statement

Despite the ubiquity and influence of online financial forums, there exists a significant gap in rigorous academic study aiming to gauge these platforms' impact on market sentiment and financial behavior. (1) These forums have shown their power not just to influence but to disrupt traditional financial markets on a monumental scale. Nonetheless, scholarly work has not kept pace in systematically examining these platforms' utility as potential bellwethers for market trends and investor behaviors. (1)

Given the increasingly critical role these online platforms play in shaping market activities, two key questions become inescapable:

1. How can algorithmic scrutiny of online financial dialogue offer a multi-layered understanding of current market sentiment and potential trends?
2. To what extent can these data-driven examinations yield insights that are actionable yet distinct from those obtained through conventional financial metrics?

The primary objective of this study is to explore these research gaps by deploying cutting-edge natural language processing techniques. The study encompasses posts and discussions culled from 11 diverse financial subreddits during a specific timeframe: late July to late August 2023. Through sentiment analysis and topic modeling, the research aims to shed light on how decentralized financial discourse may function either as a supplement or a challenge to established financial market indicators. Again, it is vital to note that the findings should not be considered financial advice but rather an exploration of sentiment trends within the limited scope of 1,200 posts.

2: Data Collection

Reddit API Description

To collect data for this study, the Python Reddit API Wrapper (PRAW) was employed. PRAW allows easy access to Reddit's API, facilitating the extraction of subreddit posts and associated comments. A Reddit application was registered to obtain the client_id, client_secret, and user_agent parameters essential for authenticating API requests. (3)

2.1: Data-fetching Limits and Handling

Reddit's API imposes rate limits on the number of requests that can be made within a specific time frame. To adhere to these restrictions, a two-second pause was introduced between each post extraction. Moreover, if the rate limit was exceeded, the program was designed to pause for 60 seconds before resuming the data fetch. Up to three retry attempts were made for each subreddit to ensure robustness in data collection. (4)

2.2: Data Description

The study concentrated on 11 selected financial subreddits ('investing_discussion', 'dividends', 'investing', 'wallstreetbets', 'StockMarket', 'DueDiligence', 'SPACs', 'ValueInvesting', 'SecurityAnalysis', 'Wallstreetbetsnew', and 'options'), with each subreddit varying in investment philosophy and financial discourse. From each subreddit, the top 150 posts were extracted based on a monthly time frame, along with up to 10 top-level comments per post. The acquired data was then stored in a Python Pandas DataFrame before being exported to a CSV file named reddit_stock_market_data_cleaned.csv.

2.3: Data Exploration

Each DataFrame entry included the subreddit name, post title, post body, and the top-level comments. The comments were stored as a list of strings within each DataFrame row. After fetching data from all the subreddits, the DataFrames were concatenated into a single DataFrame to facilitate further analysis.

3: Pre-Processing (Cleaning the comments)

3.1: Comment Cleaning

To prepare the data for analysis, several cleaning steps were executed on the comments. Given that the focus was on studying sentiment within these posts, it was crucial to remove irrelevant or noisy text.

1. Investing_Discussion Subreddit: For posts from the investing_discussion subreddit, a specific type of comment detailing a personal experience with Bitcoin scams was removed. This type of comment was identified as noise that could potentially skew the sentiment analysis.

2. Dividends and WallStreetBets Subreddits: The first comment in the list for each post fetched from the dividends and wallstreetbets subreddits was also removed. This decision was based on the observation that the first comment often contained automated or pinned messages, which were not reflective of the general sentiment.

These cleaning processes were applied directly to the DataFrame before saving the cleaned data back into a new CSV file, reddit_stock_comments_cleaned.csv.

4.1 Introduction to Sentiment Analysis Techniques Used

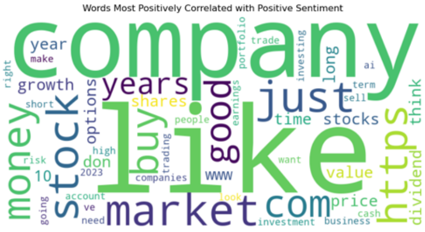
Sentiment analysis is crucial for understanding public opinion and mood regarding various stocks and investment opportunities. In this study, a multi-faceted sentiment analysis approach has been adopted. We leverage VADER (Valence Aware Dictionary and sEntiment Reasoner), TextBlob, and Natural Language Toolkit (NLTK) along with Term Frequency-Inverse Document Frequency (TF-IDF) techniques. The sentiment analysis primarily focuses on Reddit posts and comments scraped from multiple subreddits related to stock trading and investments. (5) (6) (7)

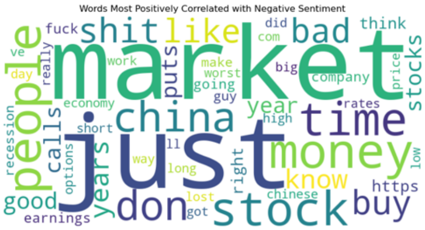
4.2: Basic Sentiment Analysis with VADER

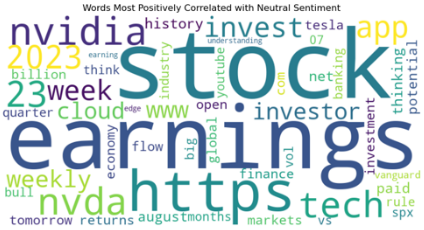
The VADER sentiment analyzer was employed to classify the sentiment of the text into three main categories: Positive, Negative, and Neutral. This classification was performed on the Title, Body, and Comments of Reddit posts. (6) Using VADER's compound score as a basis for classification, scores greater than 0 were classified as Positive, less than 0 as Negative, and equal to 0 as Neutral. VADER's lexicon is specifically calibrated for social media text, making it well-suited for analyzing Reddit discussions. (6)

Initial Findings

The model provided a quick and straightforward way to gauge the general mood surrounding various stocks and investment vehicles. Words like "like," "company," and "market" were most positively correlated with Positive sentiments, while "market," "money," and "just" were most positively correlated with Negative sentiments. This suggests that discussions about the market and companies can be polarizing, perceived either positively or negatively depending on the context and possibly the prevailing market conditions at the time of the posts. (8)


Words Most Positively Correlated with Positive Sentiment


Words Most Positively Correlated with Negative Sentiment


Words Most Positively Correlated with Neutral Sentiment

Entities and Sentiment Correlation

Positive Sentiment

Entities like 'VOO', 'SPY', 'IRA', and 'Roth' among others were most positively correlated with Positive sentiments according to VADER. This could indicate a favorable view of these investment vehicles or terms within the Reddit community. Interestingly, both 'PE' and 'EBITDA' also showed a positive correlation, albeit to a lesser extent. This provides an additional layer of support to our earlier findings where these financial metrics had moderate positive sentiment scores. (8)

Negative Sentiment

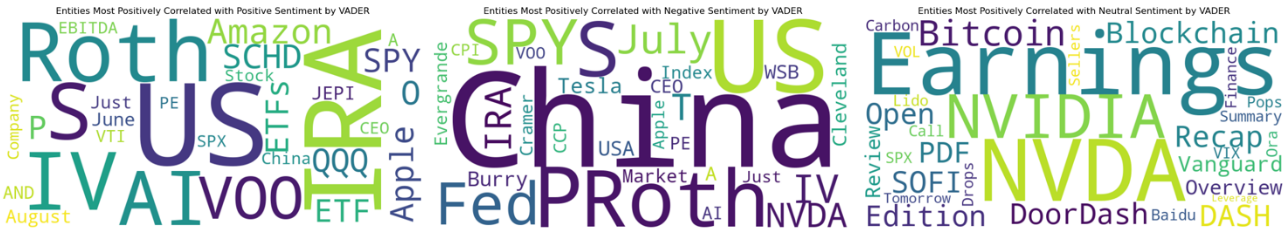
On the flip side, entities such as 'China', 'US', and 'Fed' were frequently associated with Negative sentiments. 'SPY' and 'PE' also appeared in the negative list but were outweighed by their positive mentions. This dichotomy might suggest that these topics are particularly divisive or subject to fluctuating opinions based on news or market conditions. (5) (8)

Neutral Sentiment

Entities like 'Earnings', 'NVDA', and 'NVIDIA' were most positively correlated with Neutral sentiments. The neutrality in discussions around these terms could suggest a wait-and-see attitude or a lack of strong sentiment in either direction. (5) (8)

Summary and Implications

The VADER sentiment analyzer served as an effective tool in classifying the sentiment of Reddit posts related to investing and the stock market. This methodology was particularly effective because VADER is designed to handle the nuances and slang found in social media conversations. (5) Intriguingly, certain entities showed up frequently in different sentiment categories. For instance, the entity 'US' was most positively correlated with both positive and negative sentiments, hinting at its polarizing nature in investment discussions. Words like 'IRA,' 'Roth,' and 'ETFs' were generally viewed positively, indicating a favorable outlook toward these investment vehicles. (5) Conversely, entities such as 'China' and 'Fed' were mostly found in negative contexts, reflecting prevailing investor concerns. The neutrality of terms like 'Earnings' and 'NVIDIA' implies that these topics are less emotionally charged among Reddit users. Overall, these findings not only allow for a better understanding of public sentiment but also reveal the nuanced ways in which specific entities and topics are discussed within the community. (5)


Entities Most Positively Correlated with Positive Sentiment by VADER

Entities Most Positively Correlated with Negative Sentiment by VADER

Entities Most Positively Correlated with Neutral Sentiment by VADER

4.3: Context-Based Sentiment Analysis

Even though mentions of positive and negative sentiments are helpful for simply visualizing on a word cloud, it is better to analyze these mentions in their own contexts. Therefore, here are a few comments where ‘US’ and ‘China’ were mentioned as they are the most polarizing, sentimental topics.

Analysis for the entity, "US"

Positive Contexts

Financial and Investment Wisdom: In the positive contexts, 'US' is mentioned as a market that is tracked by two ETFs, VOO and VTI. The conversation focuses on Morningstar's ratings for these two funds that track US equities. The discussion delves into the methodology used by Morningstar to rate the performance and quality of these ETFs, showing that these financial products linked to the U.S. markets are scrutinized for investment decisions. This highlights the influence and importance of U.S. equity markets in global investments.

Example: *“VOO is attempting to track the s&p 500 and VTI is the CRSP US Total Market Index.”*

Stock Market Performance: Another aspect is the past performance of US-based indexes. There is a comparison between VOO, which attempts to track the S&P 500, and VTI, which aims to track the CRSP US Total Market Index. This indicates that investors are keenly interested in how well these indices perform, underscoring the U.S. stock market's significance.

Example: *"Here are the annualized returns on each one for different time periods as of June 30th.*
VOO VTI
1 year 19.43% 18.79%
5 year 12.22% 11.28%
10 year 12.83% 12.29%"

Negative Contexts
Housing Affordability Crisis: In the negative contexts, the 'US' is discussed in the context of an increasingly unaffordable housing market. The rising home prices relative to median household incomes highlight a worsening economic situation for many American families.

Example: *“The median sales price of a home in the US is now 560% of the median household income. In 2008, it was 360% of the median household income. This is the least affordable housing market in history.”*

Economic Inequality: There's a discussion about how certain age groups and investors are faring better than others in the current U.S. housing market. This implies economic disparities and hints at issues of generational wealth and inequality within the U.S.

Example: *"I just learnt over the weekend based on Morgan Stanley's research that 33% of residential housing is owned by people of age group 60 or above in the US. And over 50% of them bought that home before the year 2000. So you're talking about a homeowner who has seen an incredible amount of home price appreciation with no mortgage attached."*

In summary, the positive contexts emphasize the importance and attractiveness of U.S. financial markets, especially for investment. The negative contexts focus on social and economic issues, specifically the affordability of housing and economic inequality.

Analysis for the entity, “China”

Positive Contexts:
General Discussion: The positive mention here doesn't seem to attribute any specific quality to China but rather brings it up as a subject for economic discussion. The mention is more neutral and invites a dialogue about the state of China's economy and its potential effects on global stock markets, particularly on QQQ type companies (tech-heavy stocks).

Example: *"But I hadn't paid much attention until recently to the state of China's economy. If there's a black swan 2008ish type event over there, what would we expect to happen with, say, QQQ type companies?"*

Negative Contexts:
Economic Woes: A strong theme that emerges is that of China's economic troubles, including slashing interest rates, deflation, and problems in the housing market. These topics contribute to a negative view of China's current state and raise questions about its future.

Examples: *"China Slashes Rates, Suspends Youth Jobless Data as Economy Signals Sharper Downturn", "China Slips Into Deflation in Warning Sign for World Economy"*

Social Contract: One mention refers to the fraying of China's social contract, which brings in social dimensions to the economic issues.

Example: *"China's Economic—and Social—Contract Is Fraying"*

Global Impact: The comments discuss the potential global ramifications of China's economic status, suggesting it could impact the world economy.

Example: *"It also wouldn't surprise me if this has a negative impact on the global economy"*

Real Estate and Unemployment: There's a specific focus on the real estate bubble and unemployment rates, including personal experiences that highlight how unaffordable property has become in China.

Example: *"I will not be able to buy a property in front tier cities of China until I'm probably in my 50s if ever."*

Media Bias: There are mentions that suggest the Western media, particularly outlets like WSJ, could be biased in their portrayal of China.

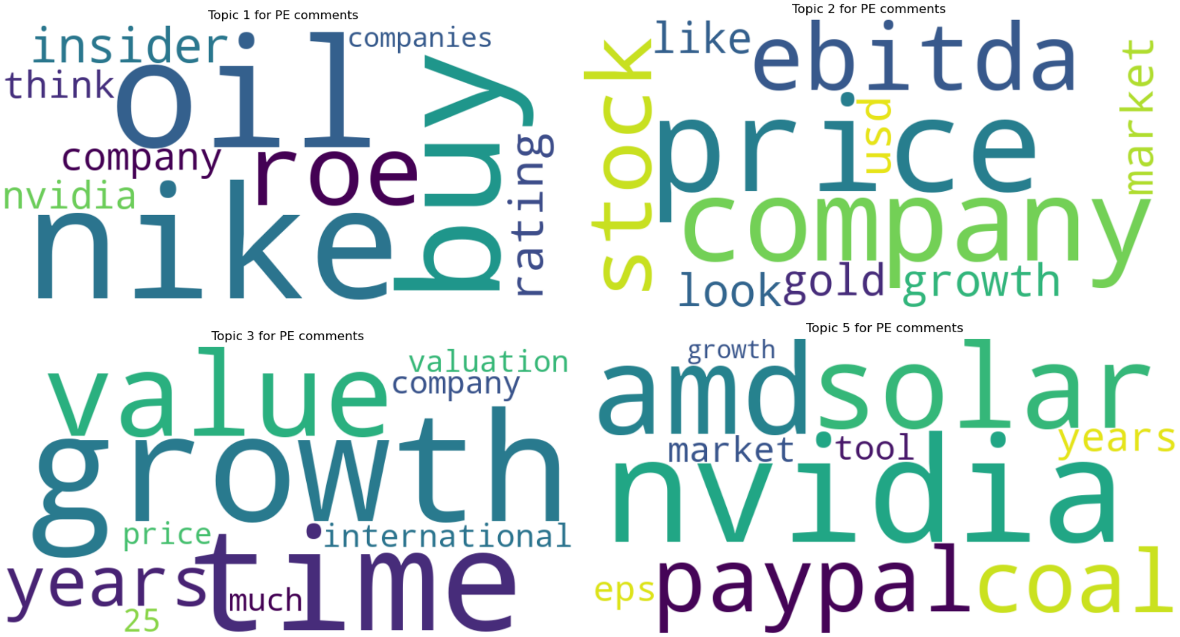
Example: *"I mean you also have to be aware that a lot of news about China we get in the west will be biased."*

Skepticism About Dire Predictions: A few comments express skepticism about the predictions of China's downfall, suggesting that such fears have been frequent but not always realized.

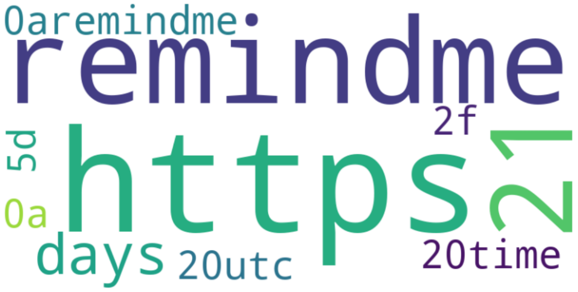
Example: *"If I had a dollar for every time someone said China was doomed, I wouldn't need to invest because I would be beyond rich already."*

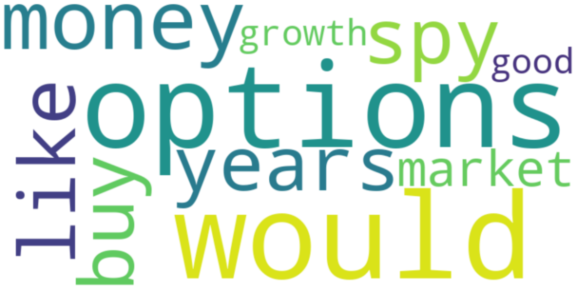
Overall, while the positive mention is more neutral and opens a dialogue, the negative mentions are more multifaceted and discuss a variety of issues, from economics to social factors, that portray China in a negative light.

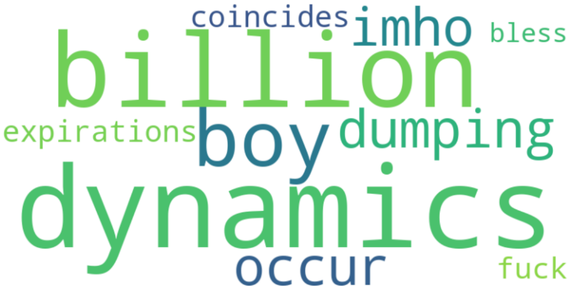
4.4: Topic Modeling Using Latent Dirichlet Allocation (LDA)
Topic modeling was performed on comments that included specific financial metrics and stocks like 'PE', 'EBITDA', 'VOO', and 'SPY'. This helped in identifying the dominant topics discussed in relation to these financial terms and popular investment options. (9) Visualizations of how sentiment distribution varies across these topics were also generated. For this report, I will only analyze the LDA distribution of PE and SPY as they yielded good results that I could use for inferring implications. (9)

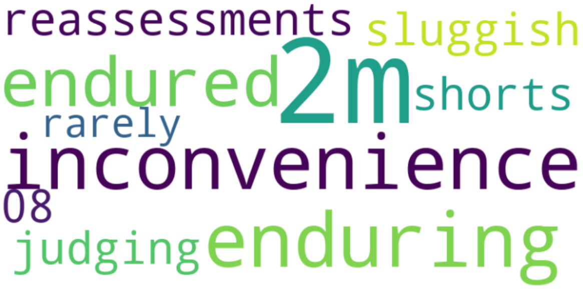

Topic 1 for PE comments
Topic 2 for PE comments
Topic 3 for PE comments
Topic 5 for PE comments

For PE (Price-to-Earnings Ratio):
Topic #2 & #5 ('company', 'see', 'cheap', 'market', 'price', 'like', 'years', 'gold', 'solar', 'growth') suggests discussions around company valuation, specifically referring to sectors like gold and solar. The mention of 'cheap' and 'growth' may indicate value investment opportunities.
Topic #3 & #1 ('valuation', 'market', 'nvidia', 'buy', 'growth', 'graham', 'oil', 'time', 'value', 'investing') appears to center around valuation methods, citing famous investing philosophies like those of Benjamin Graham, and specific sectors like oil and technology (Nvidia).


Topic 1 for SPY comments


Topic 3 for SPY comments


Topic 4 for SPY comments


Topic 5 for SPY comments

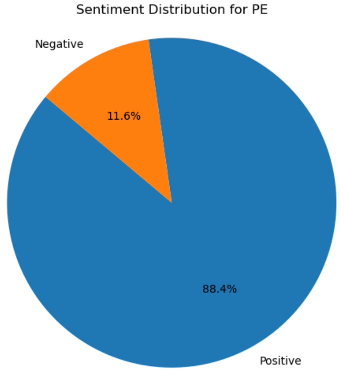
For SPY (SPDR S&P 500 ETF Trust):

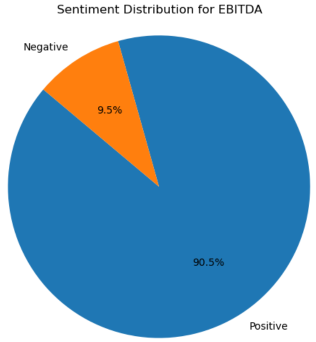
Topic #1 and #3 ('remindme', '20time', 'days', 'https', ‘options’, ‘would’, ‘buy’) suggests discussions on the right time to buy in or sell out of SPY calls or puts.

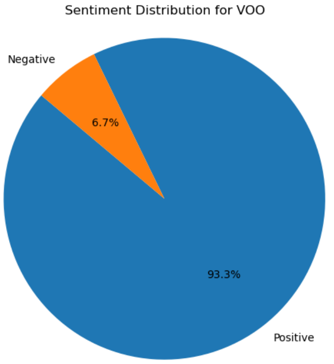
Topic #4 ('dynamics', 'billion', 'dumping') appears to center around the dynamics at play with the ETF which impact the price and the extent of the impact.

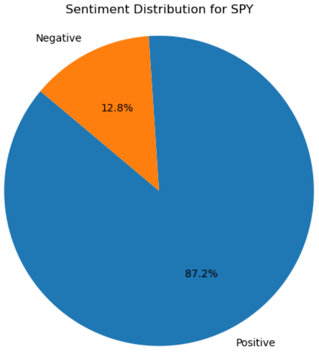
Topic #5 (‘inconvenience’, ‘enduring’, ‘judging’, ‘shorts’, ‘sluggish’) appear to be involving a negative sentiment towards a bearish view on the trajectory of SPY.

4.5 Comparative Analysis of Sentiment Scores









Comparing Financial Metrics: PE and EBITDA

1. PE (Price-to-Earnings Ratio): The sentiment score for PE is 33, indicating a generally neutral to slightly positive sentiment.

2. EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization): The sentiment score for EBITDA is 17, which is lower than PE and may signify more skepticism or caution around this metric.

Analysis: PE has a more positive sentiment than EBITDA by 16 points, suggesting that discussions around PE are generally more favorable.

Comparing ETFs and Financial Metrics: PE, VOO, and SPY

1. VOO vs PE: VOO has a sentiment score of 65, which is substantially higher than PE's 33. The data suggests that VOO is viewed much more favorably in comparison to PE, with a positive sentiment difference of 32 points.

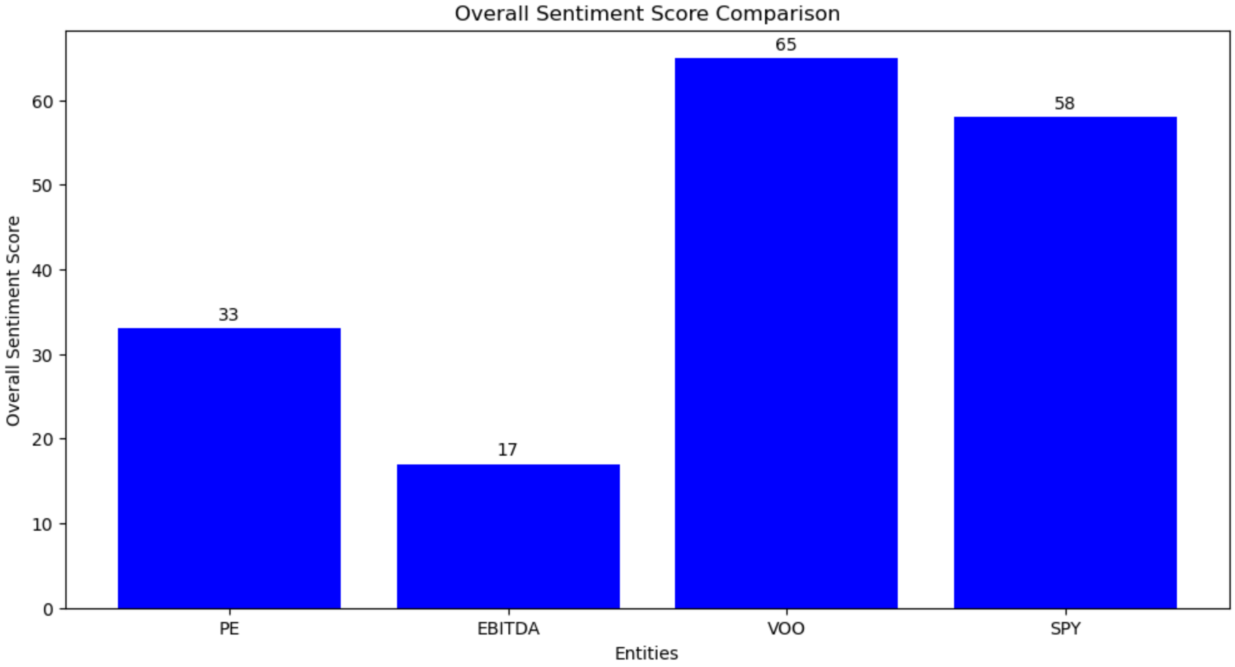
2. SPY vs PE: SPY also has a higher sentiment score of 58 compared to PE's 33, indicating that SPY is more favorably discussed by a difference of 25 points.

3. VOO vs SPY: Both are popular S&P 500 ETFs but VOO leads in sentiment with a score of 65 compared to SPY's 58. The difference is relatively minor (7 points) but suggests a slightly more positive view towards VOO.

Comparing Financial Metric and ETFs: EBITDA, VOO, and SPY

1. VOO vs EBITDA: VOO has a sentiment score of 65, which overwhelmingly surpasses EBITDA's score of 17. The data suggests that VOO is viewed much more positively, with a difference of 48 points.

2. SPY vs EBITDA: Similarly, SPY also has a much higher sentiment score (58) compared to EBITDA (17). The sentiment for SPY is more positive by 41 points.



Summary

The sentiment analysis reveals a generally more positive sentiment towards the ETFs VOO and SPY when compared to financial metrics like PE and EBITDA. Among the ETFs, VOO has a slightly more positive sentiment than SPY. Among financial metrics, PE enjoys a more favorable sentiment compared to EBITDA.

5: Conclusion:

My reddit-based sentiment analysis, employing the VADER sentiment analyzer, offers valuable insights into the Reddit community's stance on various topics related to investing and the stock market’s sentiment from July-August of 2023. The methodology was found to be highly effective in navigating the complexity and nuance of social media discourse. It surfaced distinct attitudes towards various entities and investment topics, such as the polarizing nature of the 'US,' the generally positive view of investment vehicles like 'IRA,' 'Roth,' and 'ETFs,' and prevailing concerns surrounding 'China' and the 'Fed.' (10) (11)

Detailed context-based sentiment analysis further illuminated these findings. For instance, discussions related to the 'US' oscillated between admiration for its financial markets and criticism of its socioeconomic disparities. In contrast, the negative sentiment towards 'China' was multi-layered, touching upon economic, social, and global impacts. Topic modeling added another layer of understanding, highlighting key topics under discussion like company valuation, investment timing, and market dynamics. These were particularly evident in conversations around financial metrics and popular ETFs.

The comparative sentiment scores offered a numeric representation of these attitudes. ETFs like VOO and SPY consistently outperformed financial metrics like PE and EBITDA in sentiment scores, indicating a more favorable outlook towards these investment options. Among the metrics, PE had a more favorable sentiment, perhaps suggesting a general leaning towards traditional valuation methods over more complex metrics like EBITDA. (12) (13)

These findings have several implications:

1. Investment Strategy: The prevailing positive sentiment towards ETFs like VOO and SPY suggests that these are widely considered reliable investment vehicles, potentially guiding new investors in their decision-making process. (12) (13)
2. Risk Assessment: The negative sentiment surrounding 'China' and the 'Fed' suggests areas where investors may exercise more caution. Financial institutions might consider this data when forming risk assessments. (10) (11)
3. Public Policy: Understanding the public sentiment around economic issues like housing affordability in the U.S. or economic conditions in China could serve as valuable feedback for policymakers. (10)
4. Community Engagement: Recognizing the nuances in sentiment towards different topics can help online platforms and financial news outlets tailor their content to better engage with their audience.
5. Further Research: The polarity of sentiments associated with the 'US' and the multifaceted negative sentiments towards 'China' underscore the need for additional research to understand the factors driving these opinions more clearly. (11)

In sum, this study serves as a robust foundation for understanding public sentiment on investment-related topics in the Reddit community, providing both quantifiable data and nuanced contextual understanding that could be invaluable for investors, policymakers, and researchers alike.

Bibliography

1. Chohan, U. W., & Kerckhoven, M. (n.d.). Activist Retail Investors and the Future of Financial Markets: Understanding the R/wallstreetbets Movement. Routledge. Retrieved from https://www.routledge.com/Activist-Retail-Investors-and-the-Future-of-Financial-Markets-Understanding/Chohan-Kerckhoven/p/book/9781032397252

2. The Hive Index. (n.d.). Investing Subreddits List. Retrieved from https://thehiveindex.com/topics/investing/platform/reddit/

3. PRAW Documentation. (n.d.). Reddit PRAW API. Retrieved from https://praw.readthedocs.io/en/stable/

4. Reddit Help Center. (n.d.). Reddit Data API Wiki. Retrieved from https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

5. PyPI. (n.d.). Vader Sentiment Analysis. Retrieved from https://pypi.org/project/vaderSentiment/

6. NLTK Documentation. (n.d.). Natural Language Toolkit. Retrieved from https://www.nltk.org/

7. Scikit-learn Documentation. (n.d.). Tf-IDF Vectorizer. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

8. PyPI. (n.d.). Wordcloud. Retrieved from https://pypi.org/project/wordcloud/

9. Scikit-learn Documentation. (n.d.). LDA (Latent Dirichlet Allocation). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

10. The Guardian. (2023, August 29). Evergrande shares plummet as China economy fears mount. Retrieved from https://www.theguardian.com/business/2023/aug/29/evergrande-shares-china-economy-fears

11. Federal Reserve. (n.d.). Recent Postings. Retrieved from https://www.federalreserve.gov/recentpostings.htm

12. Nasdaq. (n.d.). VOO Nasdaq Page. Retrieved from https://www.nasdaq.com/market-activity/etf/voo

13. Nasdaq. (n.d.). SPY Nasdaq Page. Retrieved from https://www.nasdaq.com/market-activity/etf/spy