The Reddit Investor's Network: Mapping User Interactions with Stocks

Thomas Rønnebek Hansen^{S200513} **and Gediminas Leisis**^{S243093}

This manuscript was compiled on January 6, 2025

Financial debates have increasingly moved to social media sites, where ordinary people share market views and investment methods. This study investigates the network structure and sentiment dynamics of Reddit's r/stocks community by conducting a thorough examination of user interactions and market sentiment trends. Using a dataset of 1,499 nodes and 20,809 edges acquired through Reddit's API, we used the Louvain algorithm to identify unique user clusters and got a modularity score of M=0.2388, and used eigenvector centrality measurements to identify critical information brokers within the network. Sentiment analysis using the FinBERT model indicated considerable differences in community responses to market events, particularly during NVIDIA's earnings announcements, when sentiment volatility ranged from -0.21 to 0.12 across communities. Our temporal investigation revealed that community sentiment acts as a leading signal of price fluctuations, particularly in the one-week period following earnings announcements. Analysis of external link-sharing patterns demonstrated a preference for well-known financial news sources, with groups exhibiting distinct information processing techniques. These findings provide new insights into retail investor behavior in digital spaces, but data collection limitations and potential sampling bias toward large-capitalization stocks highlight the need for expanded analysis that includes broader market coverage and additional investment communities.

Network Science | Reddit | Stocks |

In recent years more and more people are exchanging information, opinions and strategies in specialized domains such as financial markets via social media. One prominent example is the ”r/stocks” subreddit, a digital forum on Reddit where users discuss stock market trends, share investment strategies and analyze market movements. The importance of online financial communities became relevant during Gamestop (GME) short squeeze of early 2021, event where retail investors participated in discussions on subreddit's such as ”r/WallStreetBets” and took coordinated action against hedge funds. This event showed the power of collective sentiment and influence of online communities (1). This project seeks to investigate the dynamics of the ”r/stocks” subreddit by employing network analysis, sentiment analysis and visualization techniques to uncover patterns of interaction, community behavior, and sentiment trends.

One of the central aspects of this project is understanding the community structure within the r/stocks network. Using the Louvain algorithm for community detection, this research aims to identify communities of users who interact frequently and are likely centered around specific topics or stocks. By analyzing the sentiment within the largest communities, particularly for the top five most-discussed stocks, this project seeks to explore how sentiment varies between communities. For example, certain communities may show bullish sentiment toward specific stocks, while others may lean bearish.

The network analysis focus on examining connections between users and posts within ”r/stocks”. Users who post or comment are represented as nodes and comments to a post are represented as edges. Metrics such as in-degree and out-degree distributions, betweenness and eigenvector centrality help to identify most popular, influential users. Modularity scores will provide quantitative insights into the strength of community structure within the subreddit. Sentiment analysis will be used to classify user comments as positive and negative (bullish and bearish), and will be used to compare sentiment trends against different communities. Using TF-IDF, this project will highlight key terms driving the discussions within each community.

Significance Statement

This study broadens our understanding of retail investor behavior in online communities by examining complex network structures and sentiment patterns on Reddit's r/stocks platform. Through an in-depth examination of user interactions and market sentiment, we provide essential insights into how social media influences investment debates and decision-making. The study's creative combination of network analysis, sentiment evaluation, and community recognition exposes previously overlooked aspects of retail investor communities, notably their reactions to major market events such as earnings announcements. These findings have interesting implications for understanding today's financial markets, as social media is rapidly influencing investment behavior. The study provides useful methodological tools for assessing online investing groups, which will aid both academic research and practical market analysis.

Author affiliations: [a]Technical University of Denmark - DTU; [b]Technical University of Denmark - DTU

T.R.H was the driving force on the communities stock sentiment and financial data correlation analysis as well on the network metrics, such as degree distribution, betweeness centrality and link analysis.
G.L was the driving force on sentiment analysis, TF-IDF, wordclouds, communities sentiment and its user including the stand alone of GOOGL. BOTH, in general we were both equally involved in everything but one was the driving force. Furthermore, both was equally involved in methods, introduction,abstract, significant statement and discussion

The authors declare no competing interests.

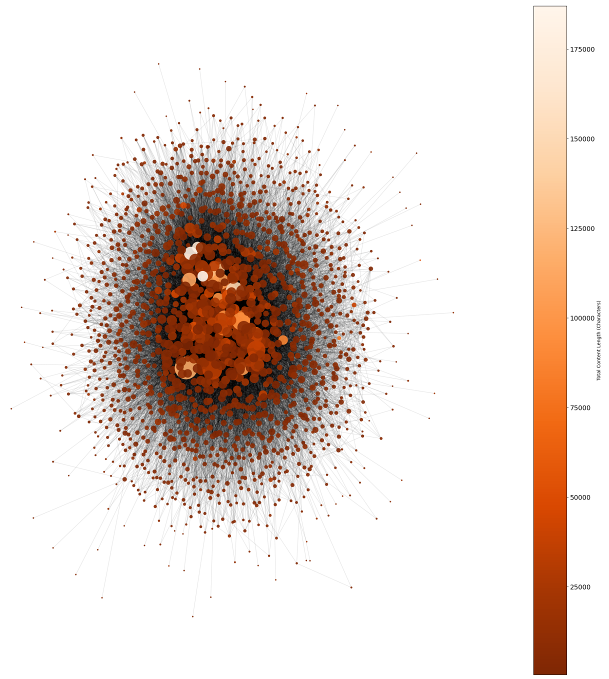

Fig. 1. Visualization of the ”r/stocks” network, with each node representing a user and edges representing user-to-user interactions. Nodes are colored based on their total content length using an inverted ”Oranges” colormap, with lighter hues representing users who write more text. Node sizes are in proportion to their degree, emphasizing more widely related persons.

1. Results

Network Structure. The ”r/stocks” network consist of 1499 nodes and 20,809 edges and all nodes and edges is contain withih the largest weakly connected component ensuring a single cohesive network structure. This indicates that the this subreddit discussion space is highly interconnected generating high amount information flow among its users.

Degree Distributions and Prominent Users. The network’s degree distribution follows a power-law distribution, showing that a small number of highly connected ”hub” users dominate the connectivity landscape. By examining in-degree and out-degree , we gain insight into user influence and engagement dynamics. As seen in Table 1, the top five nodes with the highest in-degree get significant attention, generating multiple responses and interactions from other users. In contrast, the top five individuals with the highest out-degree are significant content initiators, actively shaping the discourse through frequent posts and steering the community’s topical flow.

Centrality Measures and Influential Nodes. Beyond basic connectivity, centrality measurements could help us understand how information flows through ”r/stocks” and identify the users who influence these dynamics. Individuals with high betweenness centrality—such as ”Didntlikedefaultname,” ”dvdmovie1,” ”JRshoe1997,” ”FarrisAT,” and ”creemeeseason”—act as structural ”brokers,” connecting various sub-communities and allowing for the quick spreading of fresh ideas or breaking market news. Their positions enable them to influence the direction and content of debates, successfully connecting disparate conversational strands.

Rank	Node	In-Degree
1	ShadowLiberal	129
2	Vast_Cricket	127
3	95Daphne	115
4	FarrisAT	97
5	bartturner	90

Rank	Node	Out-Degree
1	Didntlikedefaultname	401
2	notreallydeep	278
3	creemeeseason	257
4	dvdmovie1	236
5	JRshoe1997	232

Table 1. Top five nodes with the highest in-degree (top) and out-degree (bottom).

When these centrality-based insights are combined with the degree distribution findings from Section 1 (see Table 1), certain people emerge as prominent across several metrics of impact. Looking at ”creemeeseason” and ”FarrisAT” has a high out-degree but also a significant bridging capacity. This multimodal prominence shows that certain users of the network that uses their visibility and strategic positions within the network to influence conversation.

Link Analysis and External Information Sources. To discover what type of content influential users distribute—and whether any attempts at earning money exist—we investigated the links shared by these key persons. Contrary to predictions that high-profile users would favor affiliates or sponsored content, the data reveals a preference for established financial news sources such as ”QZ.com” and ”CNBC.com” and tools for technical stock analisys like ”finviz.com.” Furthermore, individuals like ”creemeeseason” routinely link community members to conversations like ”Rate my portfolio” and ”Stocks Daily Discussion and Fundamentals,” creating an environment oriented on collaborative analysis and educated debate. Whereas ”FarrisAT” links primarily to ”i.imgur.com” and aslo ”bespokepremium.com” that is a piece of software that is used for technical analysis. This suggest that ”FarrisAT” drives the network with technical analysis of stocks and other users is asking which software he is using to make his stock observations and predictions.

These findings indicate that the network’s top users are not primarily using their positions to generate revenue through affiliate links. Instead, they tend to focus on credible, data-driven content and community-based analysis. As a result, the ”r/stocks” subreddit emerges as a more mature forum, with individuals striving for objective insights and evidence-based decision-making rather than engaging in promotional or profit-seeking conduct. This emphasis on reputable sources and analytical depth creates the assumption that information flows inside ”r/stocks” are shaped by rigorous research and communal trust rather than commercial motives.

Analysis of Stock Popularity and Sentiment. The examination of user interactions revealed significant patterns in discussion frequency and sentiment distribution. GOOGL received the most mentions (11,104), followed by AAPL (9,371), with AMD, TSLA, and NVDA in second place (6,149-6,270). GOOGL and AAPL’s prominence indicates their market leadership, whereas the secondary tier highlights interest in high-growth industries such as electric vehicles (TSLA) and semiconductors (AMD, NVDA).

Sentiment analysis revealed that neutral sentiment predominate (69.9-71.9%), indicating fact-based debates. AAPL received the most positive emotion (18.1-19.7%) as a result of its good performance, while TSLA received slightly more negative sentiment (9.5-11.0%), which is most likely related to controversy. NVDA's low negativity is consistent with its stable perception in AI and gaming.

Community analysis. To get a better idea of how sentiment can vary in different groups, we discovered the top 5 communities with most users in the "r/stocks" network using the Louvain algorithm. The modularity score of communities was calculated as $M = 0.2388$ indicating suboptimal partition(2).This suggests that "r/stocks" subreddit has noticeable clusters of user interactions, but also has significant overlap between the clusters.

Community 4		**Community 1**	
Node	**Eigenvector Centrality**	**Node**	**Eigenvector Centrality**
Vast_Cricket	0.1399	Chornobyl_Explorer	0.0688
JRshoe1997	0.0959	Invest0rnoob1	0.0665
stickman07738	0.0792	Narrow_Elk6755	0.0603
SpliTTMark	0.0789	SpongEWorTHiebOb	0.0579
Ehralur	0.0744	peter-doubt	0.0575
Community 3		**Community 0**	
Node	**Eigenvector Centrality**	**Node**	**Eigenvector Centrality**
ShadowLiberal	0.1386	95Daphne	0.1553
FarrisAT	0.1174	creemeeseason	0.1227
bartturner8	0.1071	Lost-Cabinet4843	0.1146
mayorolivia	0.1051	dvdmovie1	0.1122
notreallydeep	0.0761	WickedSensitiveCrew	0.1014
Community 2			
Node	**Eigenvector Centrality**		
Straight_Turnip7056	0.1131		
Spins13	0.0744		
skilliard7	0.0688		
TheJoker516	0.0679		
Andrew_Higginbottom	0.0515		

Table 2. Top five nodes with the highest eigenvector centrality in each community.

Influential users and most popular stocks in communities. Within each community, we identified the top 5 users based on eigenvector centrality shown in Table 2. This allows us to find the most influential users by their connections to other influential users. We perceived that all the user from Table 1 can be seen in Table, proving their influence on the network. In addition top 5 most frequently mentioned stocks in each community were identified. Revealing key points of discussion and interests within each group. These stocks align well with the stocks mentioned by influential users in their respective communities, suggesting a correlation between community interests and influential users. Also, it can be noted that stocks with the highest market capitalization, such as NVDA or APPL are frequently mentioned in different communities, indicating that large capitalization stocks are dominant topics across the network.

Sentiment analysis across communities. To compare sentiment trends in different communities, user-based average sentiment was calculated for each frequently mentioned stock, and the results were visualized in a sentiment heatmap (Fig 2). The heatmap reveals that user-based sentiment is mostly positive for popular stocks with exceptions of NVDA and AMD. These stocks are especially negative in community 3, likely reflecting critical discussions surrounding GPU chip makers. The heatmap also highlights GOOGL in community 0, which exhibits overwhelmingly positive sentiment score of 0.92. This shows clear consensus and bullishness for the stock in the community. In contrast, sentiment towards NVDA shows a lack of clear consensus across communities. Communities 0 and 2 exhibit slightly positive sentiment (0.07 and 0.12, respectively), and community 1 and 3 display mildly to strongly negative sentiment (-0.04 and -0.21, respectively). Building on these findings, the part that follows delves deeper into the differences in sentiment towards NVDA.

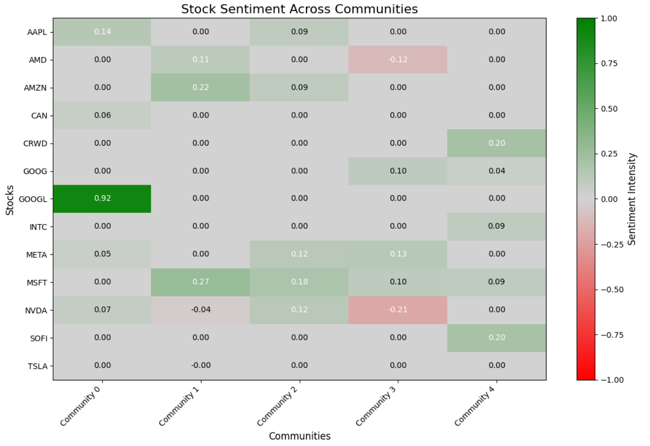

Fig. 2. Heatmap visualization of average sentiment scores across communities for frequently discussed stocks. The sentiment intensity ranges from -1.0 (deep red, indicating strongly negative sentiment) to 1.0 (deep green, indicating strongly positive sentiment), with neutral sentiment represented by gray. Values within each cell represent the mean sentiment score for a given stock within its respective community.

Analyzing community perspective on Google (GOOGL). The analysis previously highlighted hight sentiment of GOOGL (0.92) in community 0. Word cloud analysis in Fig 3 further supports this finding, highlighting frequently mentioned terms such as "good", "going", "think", "year" and "earnings". These terms suggest that discussions surrounding GOOGL in community 0 are focused on optimism about market trends, earnings calls and long term performance. The exceptional positive sentiment could be attributed to specific user dynamics or influential individuals within the community who actively advocate for the stock.

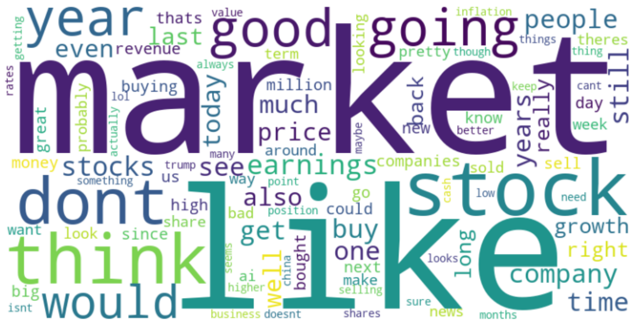

Fig. 3. The size of each word represents its relative frequency in the discourse, with "market," "think," and "stock" emerging as dominant terms. Notable market sentiment indicators include terms like "good," "growth," and "earnings," suggesting a focus on fundamental analysis

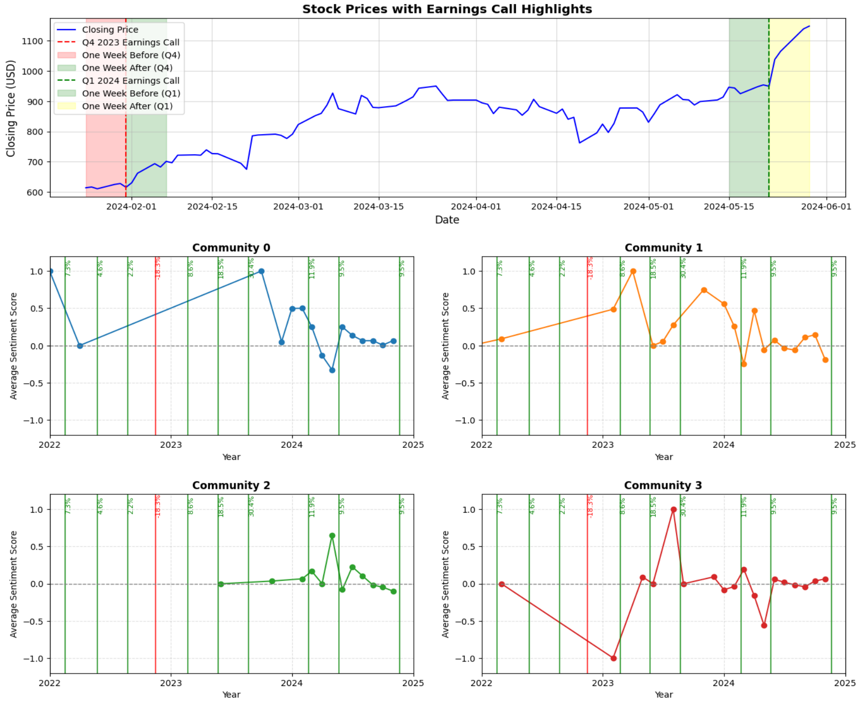

Fig. 4. The top panel shows the stock's daily closing price trajectory, with key corporate earnings events visibly identified. The vertical dashed red line represents the Q4 2023 earnings call, while the vertical dashed green line denotes the Q1 2024 earnings call. Colored shaded portions correspond to the one-week periods preceding and following each announcement, thus positioning the market's price development within well-defined pre- and post-event time frames.The lower panels show four communities, with each graph demonstrating the evolving average sentiment towards "NVDA" over a multi-year period. Vertical lines in these panels show successive earnings calls, annotated with "surprise percentage" metrics that indicate whether reported earnings beat or underperformed market expectations. Green lines show good surprises; red lines indicate negative surprises. An examination of the temporal connection between earnings announcements and community sentiment reveals varied behavioral tendencies among the communities. Following the Q4 2023 earnings call, which revealed an 11.9% positive surprise, all communities underwent an initial sentiment drop, but with various degrees of volatility. Community 0, 1, 3 had notably strong reactions, with sentiment oscillations reaching considerable amplitudes, but Communities 2 had more measured responses. This pattern occurred during the Q1 2024 results announcement, where the initial sentiment decline was followed by a convergent recovery phase that was highly connected with future stock price increase.

Exploring NVIDIA (NVDA) Sentiment Trends. The temporal analysis of community sentiment indicates distinct behavioral patterns in response to NVIDIA's earnings calls. Variations in sentiment stability and volatility resulted in community-specific traits. Community 2 had the most stable sentiment patterns, with small shifts over time and indicating a deliberate approach to evaluating NVIDIA's performance. While Community 0 remained relatively stable, it had mostly neutral opinions and restrained responses to business events. In contrast, Community 3 had significant sentiment changes, showing an increased sensitivity to company disclosures. These distinct response patterns indicate that sentiment dynamics are influenced by both external market events and community-specific behavioral traits. Variations in sentiment stability between communities reflect different methods to information processing, risk evaluation and how they interpretate the information and markets reactions. This preliminary sentiment analysis lays the groundwork for further lexical explorations.

TF-IDF and word clouds. Text frequency analysis found key linguistic characteristics in the NVIDIA community discussions. The TF-IDF scores highlighted key terminological changes, which were represented as community-specific word clouds in Fig 5. Core words such as "AI," "Market," and "Think" were consistently used throughout communities, forming a common discourse foundation based on technological and market analysis.

Sentiment-indicative words showed clear community-specific distributions. Communities 0, 1, and 2 primarily used positive market indications ("buy," "long," "good"), whereas Community 3 had a very different linguistic pattern. In this society, traditionally positive phrases arose with equal frequency as bearish signs ("sell," "short," "earnings"), particularly in contexts relating to earnings releases. These words distributions corresponded to quantitative sentiment scores. Community 3 had negative sentiment (-0.21), in contrast to positive sentiment in Communities 0 (0.07) and 2 (0.12). Community 1's balanced distribution of positive and negative terms was consistent with its near-neutral emotion score (-0.04), indicating more analytical discourse patterns.

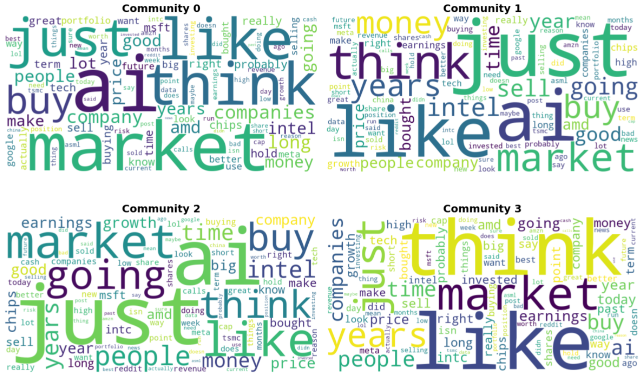

Fig. 5. Word cloud visualizations depicting key terminology across four r/stocks communities, revealing distinct discourse patterns. Word size represents frequency of occurrence, while colors differentiate semantic categories of market discussion

Discussion

Our study of retail investor behavior on Reddit reveals various methodological limits that should be considered. While our focus on the 'r/stocks' subreddit gave useful insights into retail investor discourse, the community detection showed a low modularity score of $M = 0.2388$, indicating suboptimal community partitioning.This observation, together with the exclusion of other relevant investment forums such as r/investing and r/WallStreetBets, implies that there may be limits in covering the complete range of retail investor viewpoints. These excluded communities are likely to have separate investment ideologies and debate patterns, as indicated by their diverse user bases and past responses to big market events.

This framework faced two essential limitation. First, the systematic exclusion of deleted content created a possible sample bias, particularly in divisive investment discussions. This constraint is especially apparent during times of market volatility, when contentious viewpoints may have a disproportionate influence on community opinion creation. Second, our analysis of the top 100 most discussed stocks shows a considerable bias toward large-capitalization companies. This analytical strategy, while assuring data richness, may have disguised critical insights into retail investor evaluations of potential chances for growth. Furthermore, the lack of upvote weighting in our sentiment research framework hampered our ability to quantify the relative significance of certain community debates, perhaps underestimating the impact of highly engaged content on overall investor opinion.

Expanding the dataset and addressing the limitations identified in this study could significantly enhance the depth and accuracy of insights into retail investor behavior on Reddit. By incorporating additional data sources, such as 'r/WallStreetBets' or other stock investing related subreddits, the analysis would provide more diverse perspectives and could reveal differences in sentiment across investor demographics. Recovering deleted content could offer a more complete picture of sentiment dynamics during periods of market volatility. This inclusion could capture critical, and often emotionally charged, viewpoints. Expanding our analysis to include a wider range of stocks beyond the top 100 would further reduce bias towards large-capitalization companies. This approach could help to identify talks about developing or undervalued stocks that retail investors generally overlook. Including indicators such as post and comment upvotes could reflect overall mood and interest.

Future research could broaden the investigation to additional investment-focused communities and include tools for recording controversial conversations and engagement measures. Sentiment scores weighted by user interactions may help to understand retail investor behavior across market groups and strategies.

Materials and Methods

The Reddit Investor's Network - Mapping User Interactions with Stocks

Data Collection and Processing. We gathered data from the "r/stocks" subreddit using Reddit's API via PRAW (Python Reddit API Wrapper), obtaining user interactions such as comments, posts, titles, and timestamps. To comply with API rate limits, the dataset was limited to 10,000 posts and excluded deleted content. Stock identification centered on the 100 most frequently referenced tickers during a 10-year period, checked against NASDAQ listings[3] file using regular expression matching, and resulted in a 72 MB JSON file.

Financial data - Alpha Vantage. To collect financial market data, we used the Alpha Vantage API[4] to retrieve NVIDIA's quarterly earnings releases. The data collected included reported earnings dates, quarterly EPS predictions, actual EPS values, earnings surprise percentages, and stock prices. This was used to see correlation between sentiment, stock price and earnings call.

Network Construction. We built the social network by modeling user activity in the "r/stocks" subreddit as a directed graph structure.

Users were defined as network nodes, and interactions through comments created directed edges between nodes. The comment relationship defined the direction of the edges, which directed from the commenting user to the post author or parent comment author. This solution got the hierarchical aspect of Reddit's discussion structure while maintaining interaction directionality.

Centrality Measurements. To identify influential users within the network, we utilized eigenvector centrality as primary metric. This measure evaluates user's importance within the network not only based on their direct connections but also considering influence of the users they are connected to [5]. Higher eigenvector centrality value indicates users with connections to other highly influential users, highlighting their importance within the network. Additionally, betweenness centrality was calculated to identify users who act as "bridges" connecting various sub-communities[6]. These users help to exchange information between otherwise disconnected sub-communities.

Sentiment Analysis Framework. To analyse sentiment within the 'r/stocks' subreddit, we utilized the FinBERT sentiment analysis model [7], which is pre-trained on financial communication text. This model is fine-tuned on 10,000 manually annotated (positive, negative, neutral) sentences. Using this model we calculated sentiment for every comment, every mentioned stock, and overall sentiment for each user. We applied softmax function to normalize sentiment scores, providing values that are easier to compare. The calculated sentiment was then integrated into our JSON file providing a comprehensive framework for further analysis of sentiment trends.

Temporal changes. To capture how sentiment evolved over time, particularly in response to certain events, we conducted a temporal analysis aligned with Nvidia's earnings calls. Temporal sentiment graphs were created to visualize sentiment fluctuations across different communities.

Community Detection and Analysis. We used the Louvain algorithm for community discovery and a fixed random seed (SEED = 42) to assure reliability. The algorithm's performance was assessed using modularity scoring M, which measures community strength by comparing internal link density to random network expectations.
Definition

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right],$$

where n_c is the number of communities, L is the total number of links in the network, L_c is the total number of links within community C_c, and k_c is the total degree of the nodes in community C_c.in community C_c[2]

1. A Machavarapu, Reddit sentiments effects on stock market prices (2022) Accessed: 2024-12-11.
2. AL Barabási, *Network Science.* (Cambridge University Press), p. section 9.4 (2016) Accessed online: 2024-12-11.
3. DataHub, Nasdaq stock listings (2024) Accessed: 2024-12-11.
4. Alpha Vantage, Alpha vantage api (https://www.alphavantage.co/) (2024) Accessed: 2024-12-11.
5. P Bonacich, *Social Networks.* (University of California at Los Angeles), pp. 555–564 (2007) Accessed online: 2024-12-11.
6. LC Freeman, *Sociometry.* (American Sociological Association), pp. 35–41 (1977) Accessed online: 2024-12-11.
7. HW Huang, Allen H., Y Yang, Finbert: A large language model for extracting information from financial text (2022) Accessed: 2024-12-11.