University of Glasgow

Web Science:

Assessed

Coursework

Reddit Data Analysis M-LEVEL

Author:

ATHARV JAIN

`

Indexing

1. Introduction

Reddit is one of the most influential online discussion platforms, where users actively engage in conversations across diverse topics. The InvestmentClub subreddit serves as a dynamic hub for investment discussions, attracting users interested in stock markets, cryptocurrency, mutual funds, and other financial instruments. Analyzing discussions within this subreddit provides valuable insights into engagement patterns, influential contributors, and emerging investment trends.

This study focuses on analyzing two key JSON datasets—Submissions (posts) and Comments—to understand how discussions evolve, how users interact, and which topics dominate the conversation. By leveraging network analysis, sentiment detection, and machine learning techniques, this report aims to identify super users, track discussion trends, and predict future influential contributors.

The significance of this analysis lies in its ability to uncover how communities form, how certain users drive engagement, and how investment discussions shape online financial discourse. Understanding user behavior within InvestmentClub can help investors, financial analysts, and data scientists grasp market sentiments, detect emerging investment themes, and study engagement patterns in digital financial communities.

This research not only examines historical interactions but also applies machine learning to predict future super users, thereby offering a forward-looking perspective on how subreddit engagement is likely to evolve. The findings contribute to broader discussions on social media influence in finance, community-driven investment strategies, and the role of digital platforms in shaping public market opinions.

2. Data Processing and Cleaning Steps

2.1. Identified Issues in the Dataset

Before processing the data, several issues were identified that required cleaning and transformation:

1. **Missing Values**
    - Several columns contained missing values, with some having nearly 50% or more null entries.
    - num_reports, banned_by, and removal_reason contained only null values and were unnecessary.

- author_created_utc and subreddit_type contained significant missing data.

2. **Redundant or Unnecessary Columns**
   - retrieved_on, domain, and url were metadata fields that do not contribute to engagement analysis.
   - Columns such as id and permalink were removed as they primarily serve as identifiers for external referencing rather than aiding in interaction or sentiment analysis.
3. **Standardization Needs**
   - created_utc was stored as a string and needed conversion to a datetime format for time-based analysis.
   - author_premium, locked, and quarantined were categorical but stored as objects.

2.2. Retained Fields

The following columns were retained to support network analysis, sentiment detection, and engagement tracking:

1. **Text-Based Attributes**
   - body - Content of the comment or post.
   - title - Post title (for submissions).
2. **User Information**
   - author - Username of the poster or commenter.
   - author_premium - Indicates whether the user is a premium Reddit member.
   - author_cakeday - Whether the user account was created on Reddit's anniversary.
3. **Engagement Metrics**
   - score - Net upvotes (score = upvotes - downvotes).
   - ups - Total upvotes received.
   - downs - Total downvotes received.
   - total_awards_received - Awards given by other users.
4. **Community Information**
   - subreddit - Name of the subreddit (all records belong to InvestmentClub).
   - subreddit_type - Type of subreddit (public/private/restricted).
   - locked - Whether the discussion was locked by moderators.
   - quarantined - Whether the subreddit was quarantined due to rule violations.

5. **Temporal Data**
    - created_utc - Timestamp when the comment or post was created (converted to datetime).
    - author_created_utc - Timestamp when the author created their Reddit account.
6. **Hierarchy Tracking**
    - parent_id - ID of the parent post or comment.
    - unique_parent_id - Clean version of parent_id for merging.

2.3. Data Cleaning Steps Applied

1. **Dropped Unnecessary Columns**
    - num_reports, banned_by, and removal_reason were dropped due to containing only null values.
    - retrieved_on, domain, and url were removed as they were not relevant to engagement analysis.
    - upvote_ratio and view_count were excluded due to inconsistent availability across records.
    - link_flair_text and category were removed as they did not contribute to network, sentiment, or engagement analysis.
2. **Handled Missing Values**
    - Categorical columns (subreddit_type, author_premium, locked, quarantined) were filled with 'Unknown' to maintain data consistency.
    - Numerical fields (total_awards_received) were filled with 0, assuming missing data means no awards were received.
    - Author fields (author_created_utc) were retained, but missing values were noted for possible exclusion in modeling.
3. **Converted Datatypes**
    - created_utc and author_created_utc were converted to datetime format for time-based analysis.
    - locked and quarantined were changed to Boolean values (True/False).
4. **Ensured Data Integrity**
    - Verified no unintended data loss occurred during filtering.
    - Checked for duplicate records and removed any found.

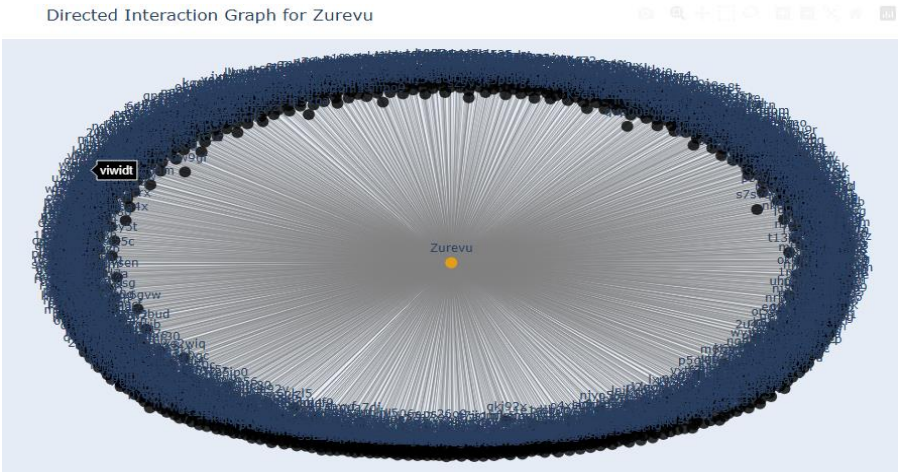
- Ensured parent_id values properly reference either submissions (t3_) or comments (t1_).

3. Graph Visualizations and Insights

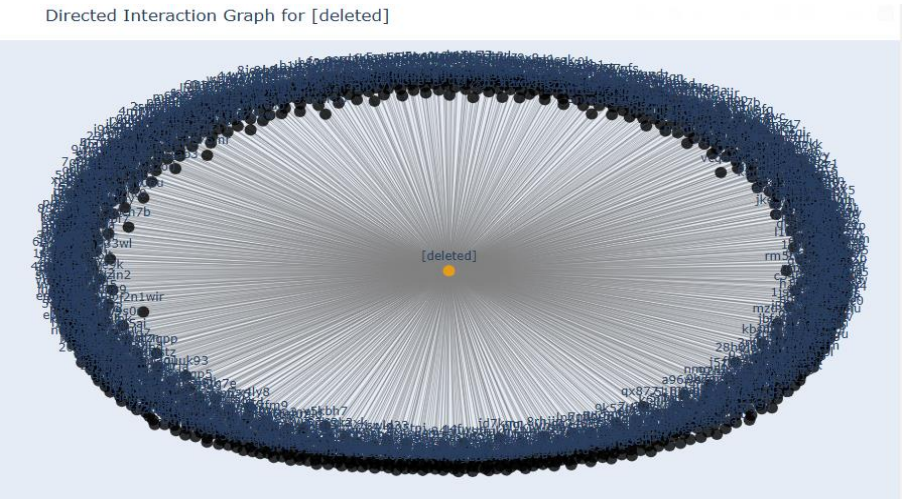
3.1 Directed Interaction Graph for the Top 5 Most Active Users

1. **Visualization Details:**
   - This graph represents interactions among the top 5 most active users, where nodes are users and directed edges indicate replies. The spring layout improves clarity by reducing overlap, while an interactive display allows better analysis. Each user's network is visualized separately, highlighting their engagement patterns within subreddit discussions.

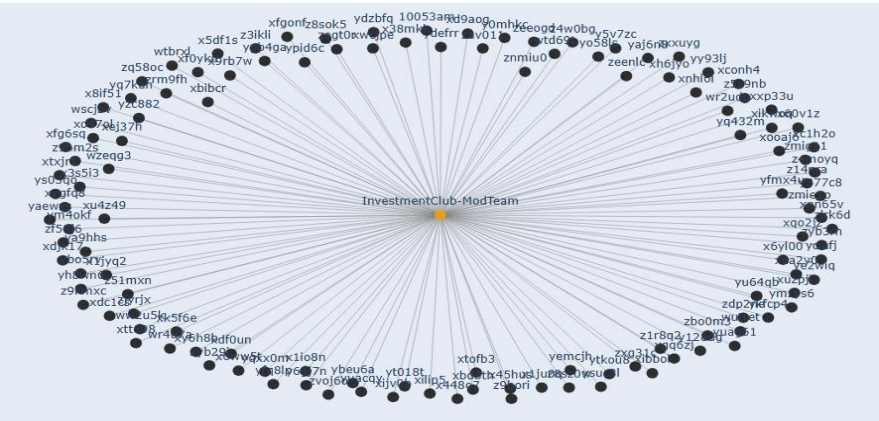
2. **Interpretation:**
   - The graph shows how key users act as discussion hubs, with some forming dense interaction networks, while others engage in more isolated conversations. Users with many edges are likely highly influential, shaping discussions. A hierarchical structure suggests some users initiate conversations, while others primarily respond, highlighting different engagement styles.



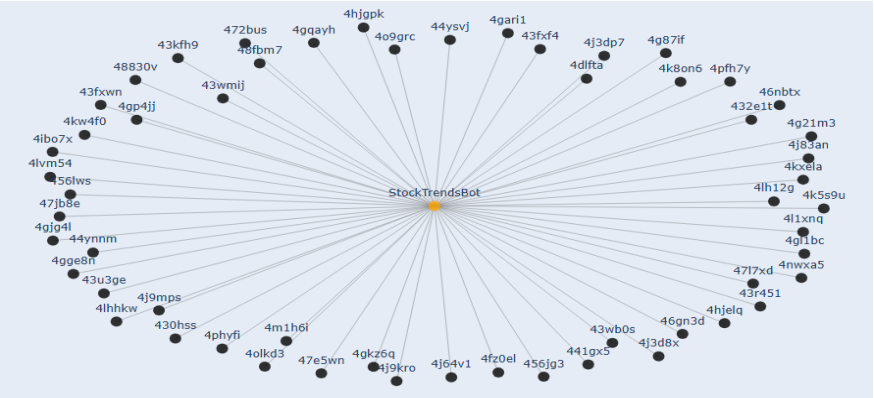
Directed Interaction Graph for [deleted]

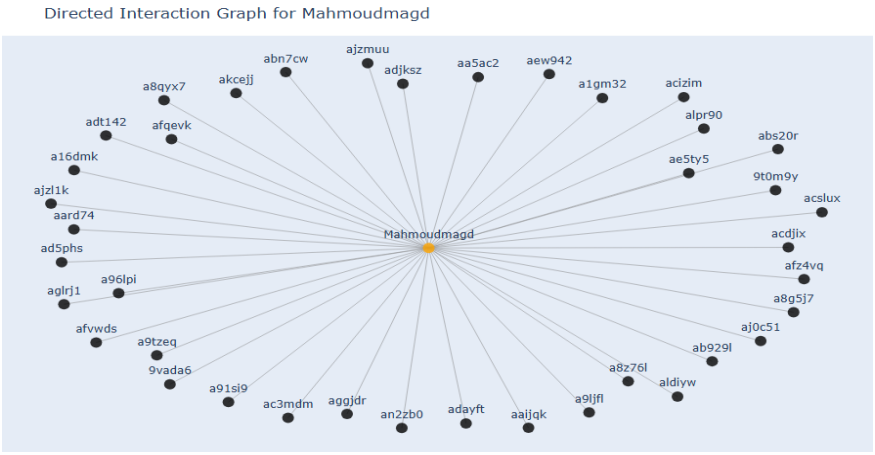


Directed Interaction Graph for InvestmentClub-ModTeam



Directed Interaction Graph for StockTrendsBot





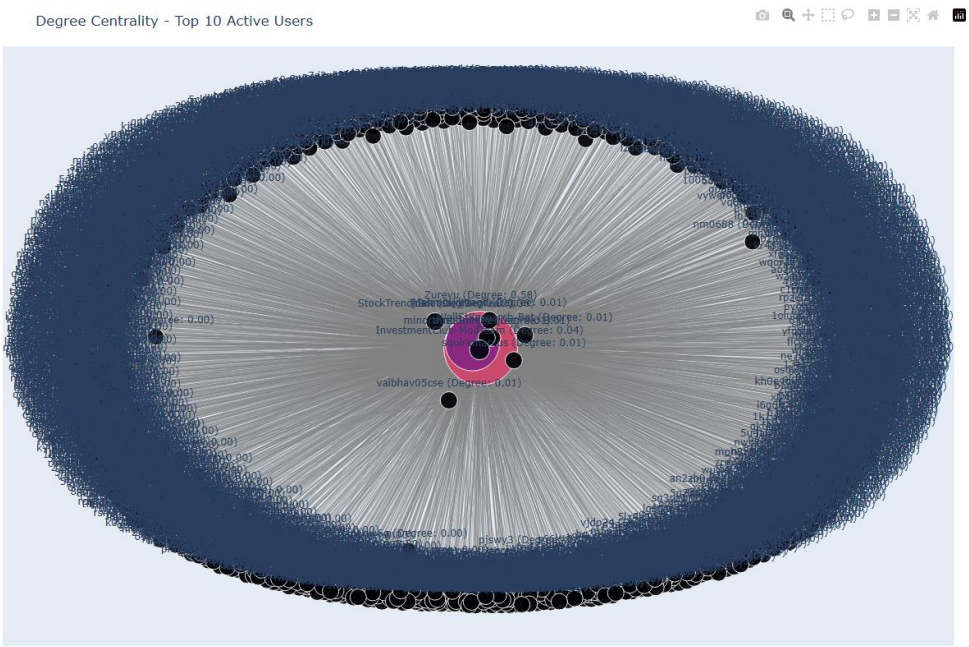
3.2 Degree Centrality Graph

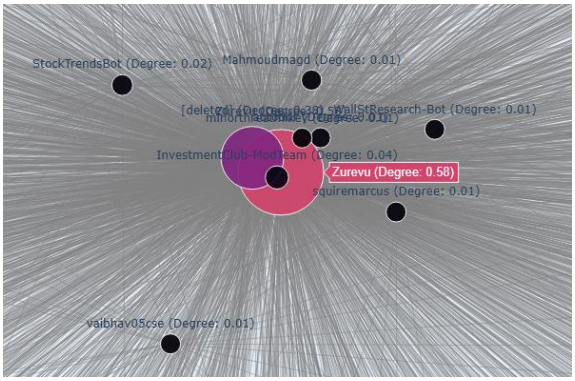
1. **Visualization Details:**
   - This graph represents the degree centrality of users, where node size is proportional to the number of direct connections a user has. A darker color scheme (Inferno colormap) is used to highlight users with higher centrality. Larger nodes indicate users with more interactions, while the spring layout ensures a structured, readable representation.

Pseudo Code –

Compute Degree Centrality Measures

degree_centrality = nx.degree_centrality(G_top_10_users)



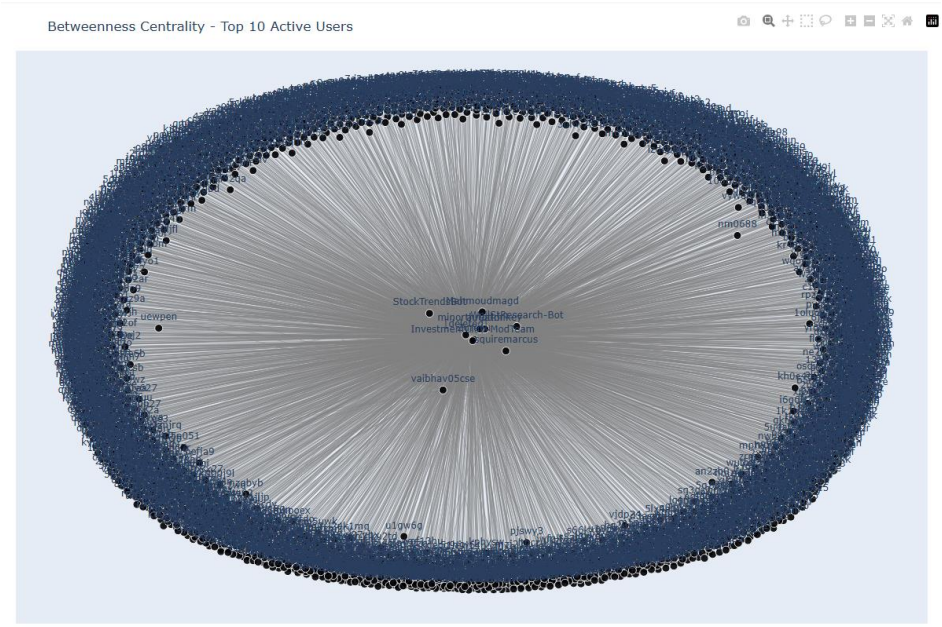


2. **Interpretation:**
   - Users with high degree centrality are highly active in discussions, frequently interacting with multiple participants. These users are key contributors, influencing engagement dynamics within the subreddit. A larger node suggests a user with significant interaction levels, shaping the overall discussion flow.

3.3 Betweenness Centrality Graph

1. **Visualization Details:**
   - This graph visualizes betweenness centrality, indicating how often a user acts as a bridge between different discussion groups. Users with higher betweenness centrality are assigned larger node sizes, and a coolwarm color scheme highlights their importance in connecting conversations.

Pseudo code- betweenness_centrality = nx.betweenness_centrality(G_top_10_users)

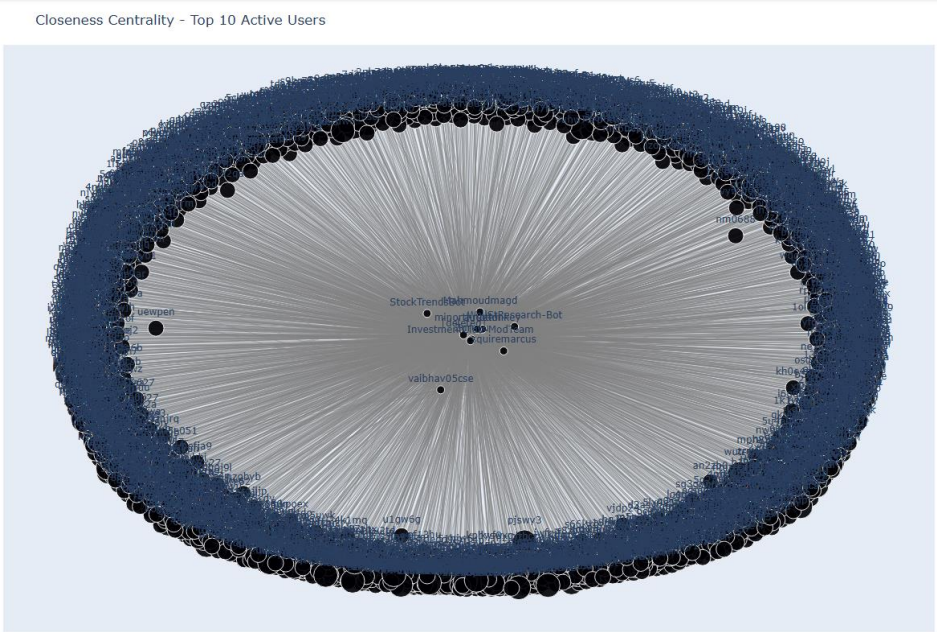


2. **Interpretation:**
    - High betweenness centrality users serve as network connectors, facilitating discussions between separate groups. These users help spread information, ensuring subreddit cohesion. If removed, conversation flow might be disrupted, demonstrating their importance in subreddit engagement.

3.4 Closeness Centrality Graph

1. **Visualization Details:**
    - The closeness centrality graph represents how quickly a user can reach others within the discussion network. The coolwarm color map and scaled node sizes indicate users who have better access to various discussion threads. Higher closeness centrality means more efficient information spread.

Pseudo Code - closeness_centrality = nx.closeness_centrality(G_top_10_users)



2. Interpretation:
    - Users with high closeness centrality are well-positioned within the subreddit to spread discussions quickly. They interact with diverse users, making them influential in discussion flow. These users ensure broader engagement, helping maintain subreddit activity and information exchange.

4. Super User Analysis

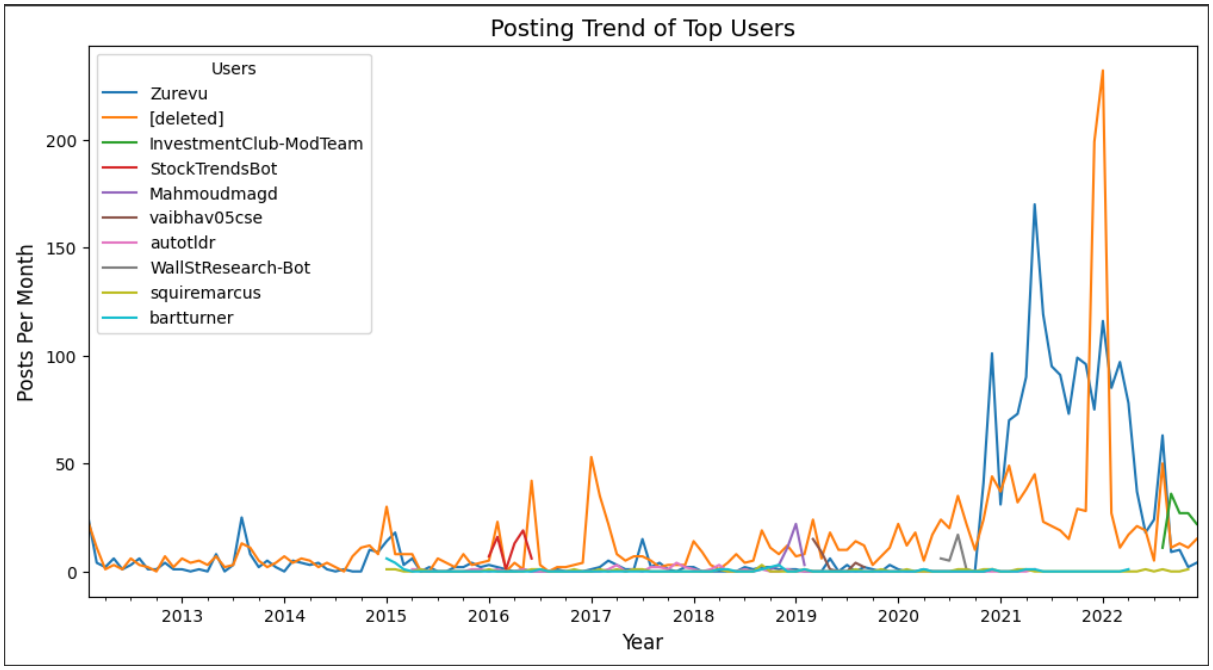
- The Super User Analysis focuses on identifying key contributors in the InvestmentClub subreddit, understanding their engagement patterns, and analyzing their influence on different investment-related topics. This analysis examines participation frequency, dominant discussion topics, network structures, and sentiment trends to understand how discussions evolve over time. This analysis aims to uncover how super users shape discussions, how communities form, and how their influence impacts engagement within the subreddit.

4.1 Identifying Super Users

- To identify super users, I calculated key metrics such as the number of posts per user, active days, posting frequency, and the time of their first and last post. These metrics allowed me to rank users based on their overall engagement levels. Additionally, I computed a posts-per-day ratio to assess how consistently each user contributed to discussions. This approach helped pinpoint the most active and influential members in the subreddit, distinguishing them from casual contributors.

author	total_posts	first_post	last_post	active_days	posts_per_day
Zurevu	2040	2012-02-01 20:20:24	2022-12-31 20:25:52	3987	0.511663
[deleted]	1963	2012-02-01 21:39:38	2022-12-30 12:48:10	3985	0.492597
InvestmentClub-ModTeam	123	2022-08-17 23:01:26	2022-12-31 22:24:49	136	0.904412
StockTrendsBot	62	2016-01-28 21:56:18	2016-06-28 07:00:56	152	0.407895
Mahmoudmagd	40	2018-10-31 20:31:48	2019-02-04 16:34:24	96	0.416667
vaibhav05cse	31	2019-03-02 14:15:43	2019-09-21 11:17:03	203	0.152709

autotldr	31	2015-04-11 23:31:29	2021-05-18 17:03:26	2229	0.013908
WallStResearch-Bot	30	2020-06-09 00:39:00	2020-12-12 20:09:10	187	0.160428
squiremarcus	28	2015-01-28 21:12:39	2022-11-09 04:05:29	2842	0.009852
bartturner	26	2015-01-09 13:12:53	2022-04-25 06:47:59	2663	0.009763



4.2 Topic Modelling and Discussion Trends

- To analyze discussion trends, I applied Latent Dirichlet Allocation (LDA) for topic modeling. First, I preprocessed the text data by removing stopwords, lemmatizing words, and applying TF-IDF vectorization to ensure a structured representation of the content. After training the LDA model, I identified the top 10 topics discussed in the subreddit. Each topic was assigned a meaningful label based on its most frequently occurring words, including "Cryptocurrency & Investments," "Stock Portfolio & Analysis," and "Mutual Funds & General Investing."

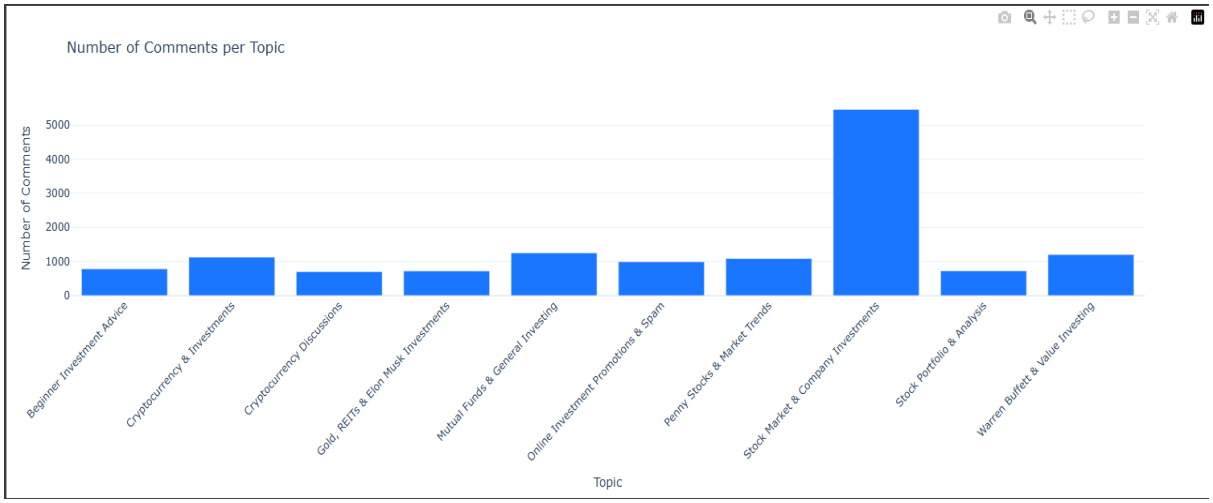
- To examine engagement patterns across these topics, I generated a bar chart representing the number of comments per topic. This visualization helped identify which investment topics attracted the most discussions and which ones were less frequently discussed. Additionally, I tracked the evolution of topic discussions over time by analyzing yearly and monthly posting trends. This enabled me to assess how different investment topics gained or lost popularity in response to market trends and events.

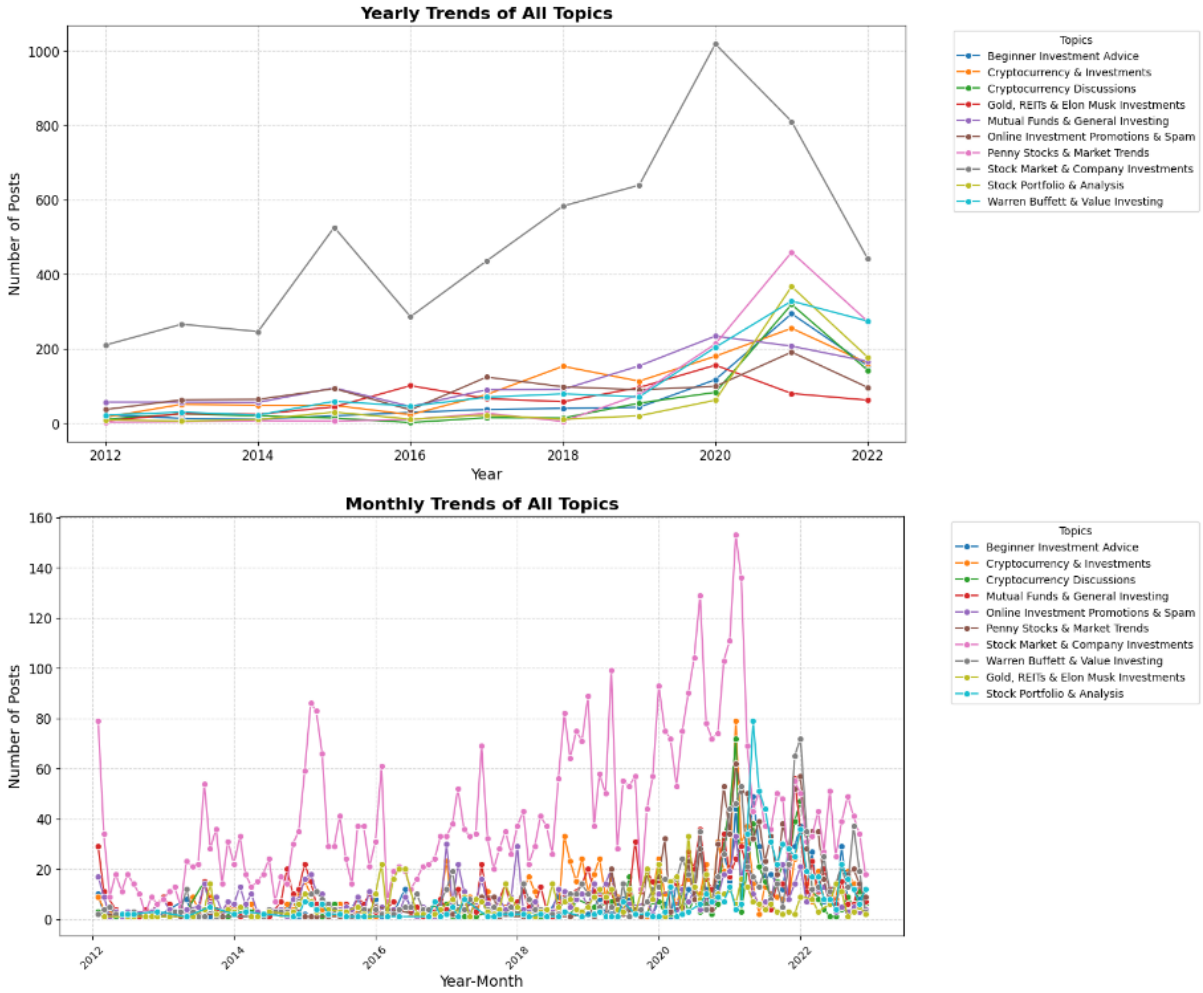
Pseudo Code-

Apply TF-IDF vectorization initialize

TF-IDF vectorizer with max features = 5000, custom stopwords, and n-gram range (1,2) transform cleaned text data into numerical TF-IDF matrix

Apply LDA Topic Modeling initialize LDA model with 10 topics and a fixed random state fit the model on the TF-IDF transformed data



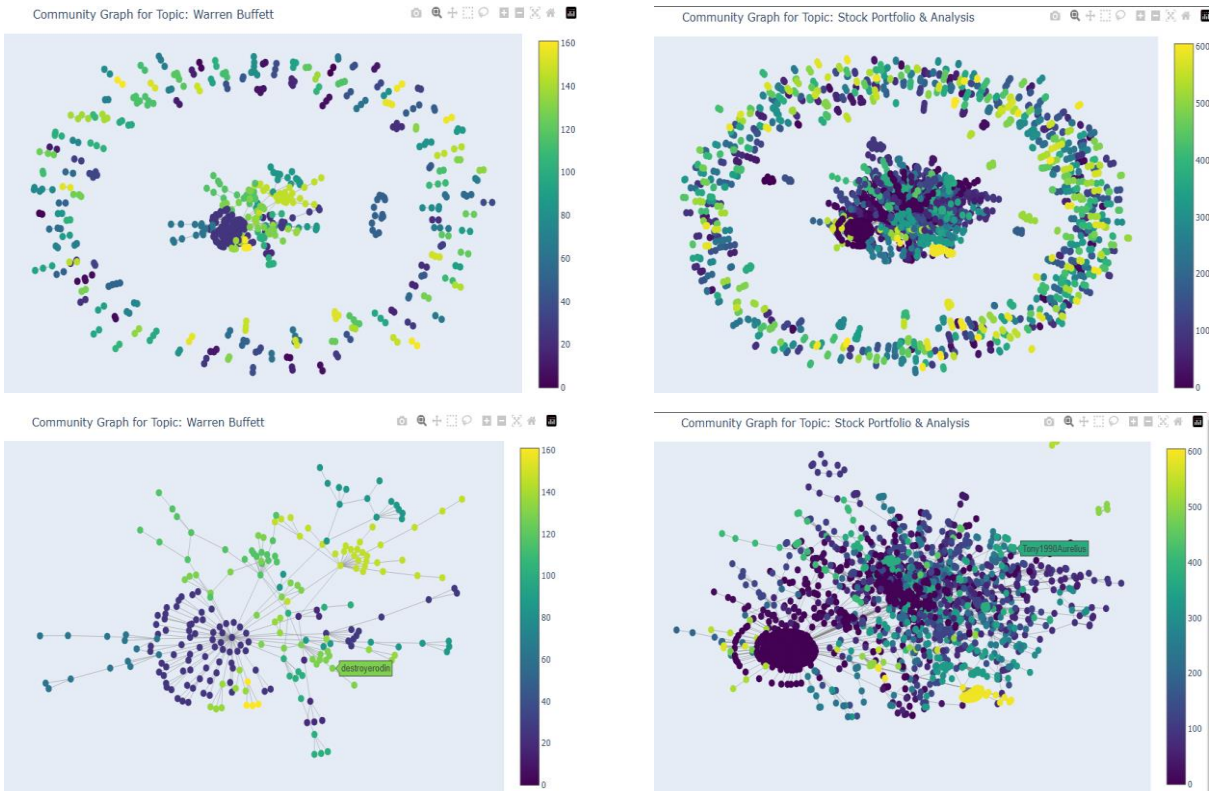


4.3 Network Graphs for Topic-Based Community Formation

- To investigate how discussions were structured, I built network graphs for each investment topic. Each graph represents users as nodes and interactions (replies) as edges to visualize how participants engage with each other. By applying Louvain community detection, I identified clusters of users who frequently interacted. This analysis provided insights into how communities form around specific investment topics and which users play central roles in sustaining discussions.

Pseudo Code -

Build and visualize community graph for each topic for each topic in dataset, create a graph with users as nodes and replies as edges visualize the graph with community detection and color-coded nodes

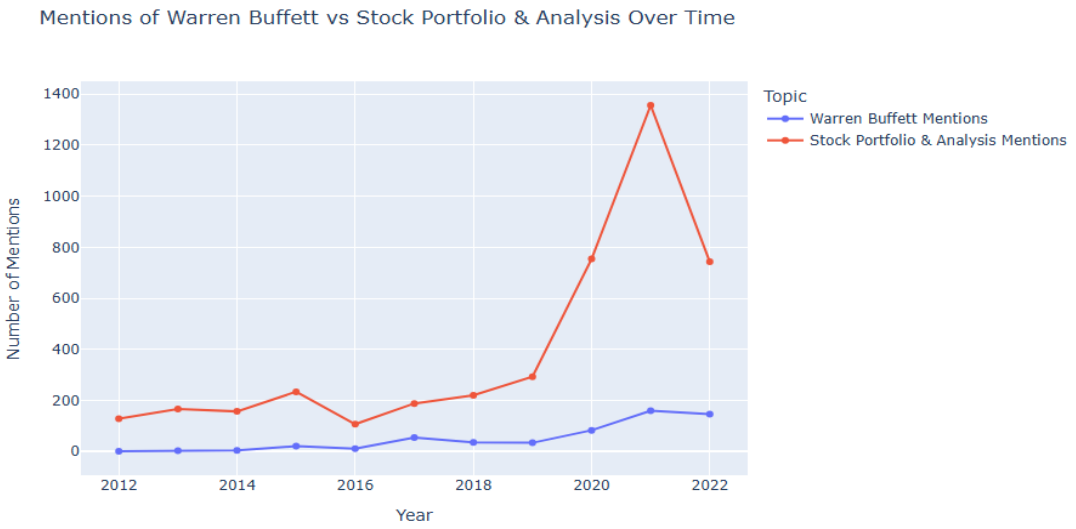


- The Stock Portfolio & Analysis graphs(right) show a dense and well-connected network, where discussions appear to be more centralized with a high number of interactions around key users.
- In contrast, the Warren Buffett graphs(left) exhibit a more dispersed structure, with some loosely connected nodes and fewer dense clusters. This suggests that discussions related to Warren Buffett might be led by a few key users with less interconnected participation from the wider community.

4.4 Comparative Analysis of Two Topics

- For a deeper analysis, I selected two widely discussed topics: "Warren Buffett" and "Stock Portfolio & Analysis". The selection was based on their high engagement levels and distinct investment themes—one focusing on a legendary investor and the other on general stock market strategies.
- I plotted a time-series graph to analyze the frequency of mentions per year, tracking how discussions around these topics evolved. This analysis helped me determine

whether these topics maintained steady engagement or fluctuated in response to market trends, economic policies, or major financial events.



- The **Stock Portfolio & Analysis** topic has significantly **higher mention volume** compared to **Warren Buffett** throughout the years. This suggests that general stock market strategies and portfolio discussions **attract more consistent engagement** within the subreddit.

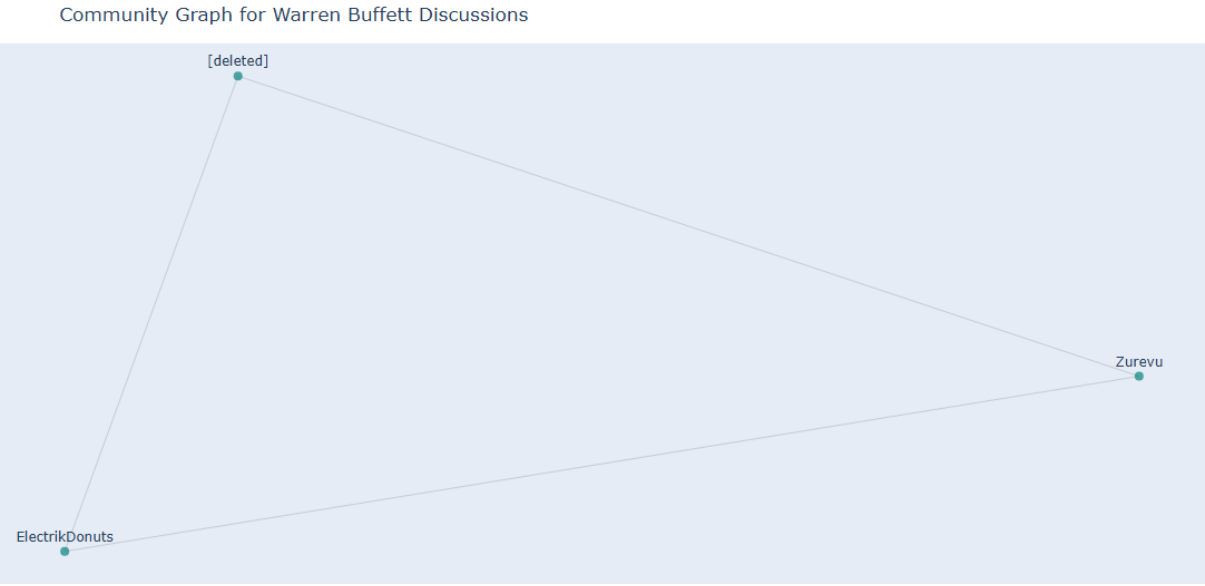
4.5 Community Detection in Super User Discussions

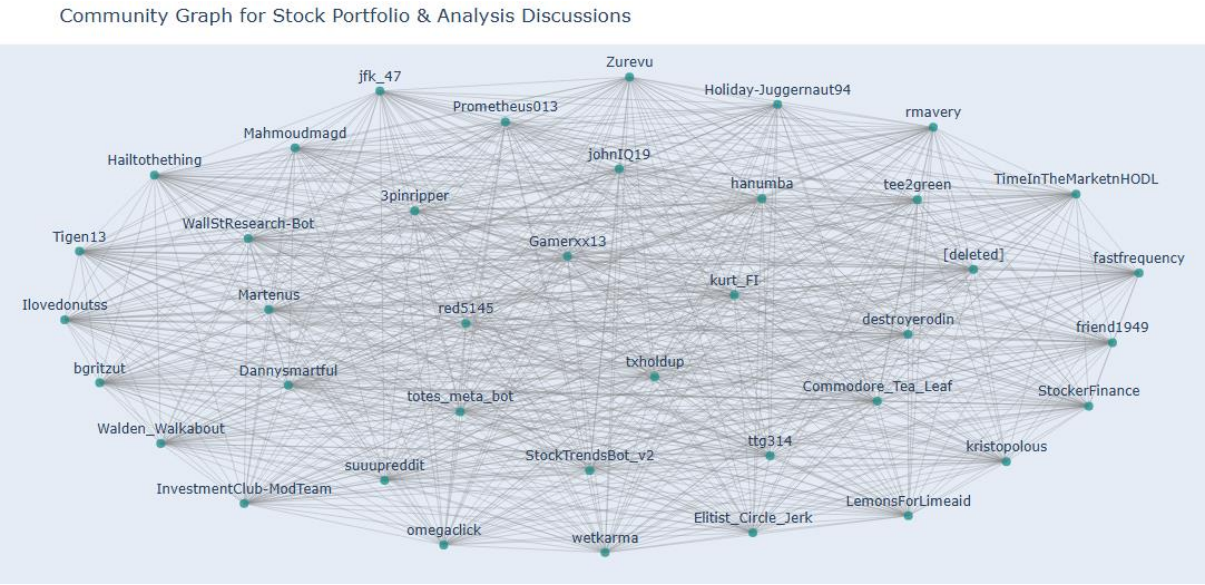
- To examine user interactions within key topics, I built network graphs where nodes represent users and edges indicate replies. Using Louvain community detection, I identified distinct user clusters in Warren Buffett and Stock Portfolio & Analysis discussions. A spring layout optimized node positioning, making patterns of engagement clearer. To refine insights, I filtered out users with fewer than five interactions and weighted edges based on reply frequency. These graphs highlight how super users influence discussions, revealing whether engagement is centralized around a few key users or distributed among a broader community.

Pseudo code –

Build Graph for Each Topic For each topic in dataset: Create an empty graph Add users as nodes Connect users based on interactions (edges)

Detect Communities and Plot Graph For each topic graph: Detect user communities Assign positions for visualization Plot interactive network graph with nodes and edges





- The Warren Buffett discussion graph has only 3 nodes and 3 edges, showing low engagement and discussions dominated by a few users with minimal interaction.
- In contrast, the Stock Portfolio & Analysis graph has 39 nodes and 741 edges, indicating high engagement with a larger, more interactive community actively exchanging ideas.

4.6 Rich Club Coefficient and Super User Influence

- To understand whether highly connected users formed an elite discussion group, I computed the Rich Club Coefficient for both topics. This metric quantifies whether high-degree nodes (super users) interact more frequently with each other than with less active users. A high Rich Club Coefficient suggests that a core group of influential users

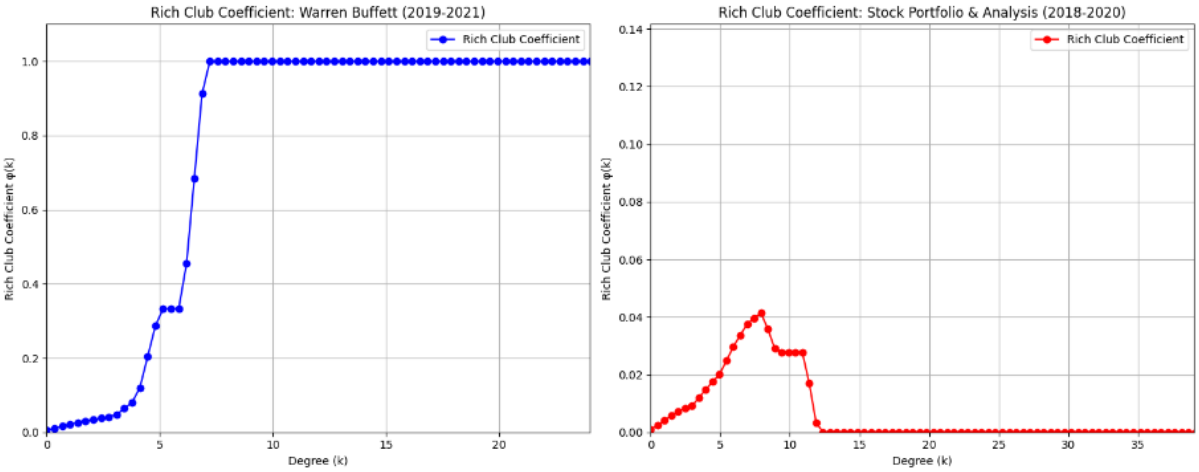
dominates discussions, while a lower coefficient indicates that discussions are more inclusive and distributed among general participants.

Pseudo Code –

Compute the Rich Club Coefficient for all degrees

rich_club_topic1 = nx.rich_club_coefficient(G_topic1, normalized=False)

rich_club_topic2 = nx.rich_club_coefficient(G_topic2, normalized=False)

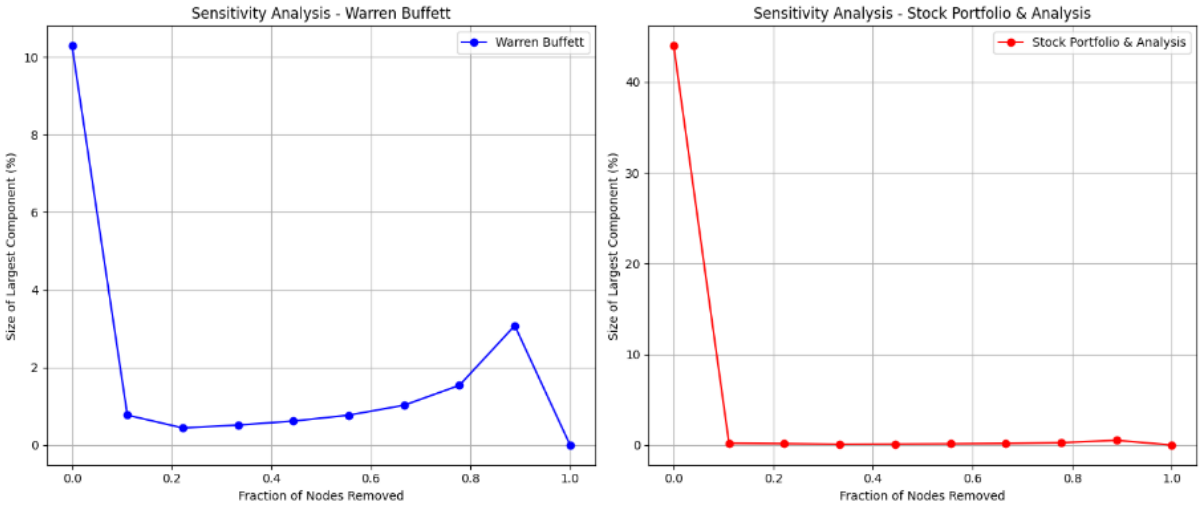


- Warren Buffett discussions have a high Rich Club Coefficient, meaning super users mostly interact among themselves, creating a centralized and exclusive network.
- In contrast, Stock Portfolio & Analysis discussions have a lower and fluctuating coefficient, indicating broader user participation rather than dominance by a small group.
- This suggests that Warren Buffett discussions are more controlled by key influencers, while Stock Portfolio & Analysis discussions foster a more inclusive and community-driven engagement.

4.7 Sensitivity Analysis for Network Robustness

- To evaluate the structural importance of super users, I conducted sensitivity analysis by gradually removing top users from the network and measuring the impact on connectivity. The key metric here was the largest connected component size, which indicates how well the network remains intact as key users are removed. The results

were plotted to illustrate whether discussions were highly dependent on super users or whether engagement remained stable even when they were removed.



- Warren Buffett discussions show gradual decline in connectivity when super users are removed, indicating a more resilient and self-sustaining network.
- Stock Portfolio & Analysis discussions collapse quickly, showing high dependency on super users, where engagement is driven by a few key contributors.
- This suggests Warren Buffett discussions are more distributed, while Stock Portfolio & Analysis relies heavily on influential users, making it more fragile.

4.8 Z-Score Analysis for User Activity

- To further investigate super user engagement, I computed Z-scores for each user based on their posting frequency. This allowed me to identify users whose posting activity deviated significantly from the average engagement level.
- I then visualized this data using scatter plots that display the relationship between the number of messages posted by a user and their Z-score. A higher Z-score indicates users with a significantly greater-than-average posting frequency, helping to highlight the most active and influential contributors.

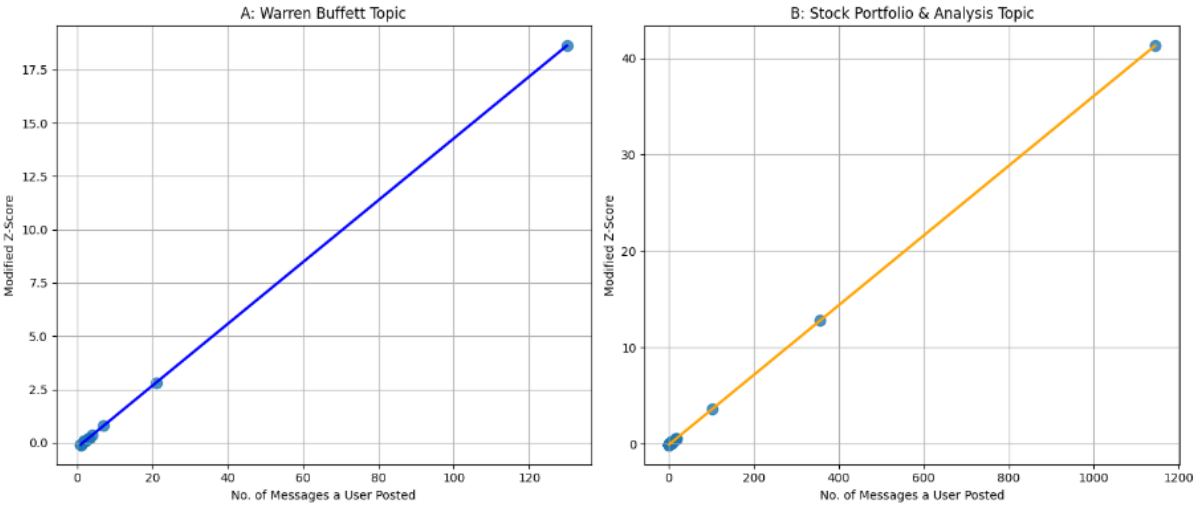
Pseudo code-

Compute Z-score for "post_count" in user_activity_topic1

user_activity_topic1["z_score"] = stats.zscore(user_activity_topic1["post_count"])

Compute Z-score for "post_count" in user_activity_topic2

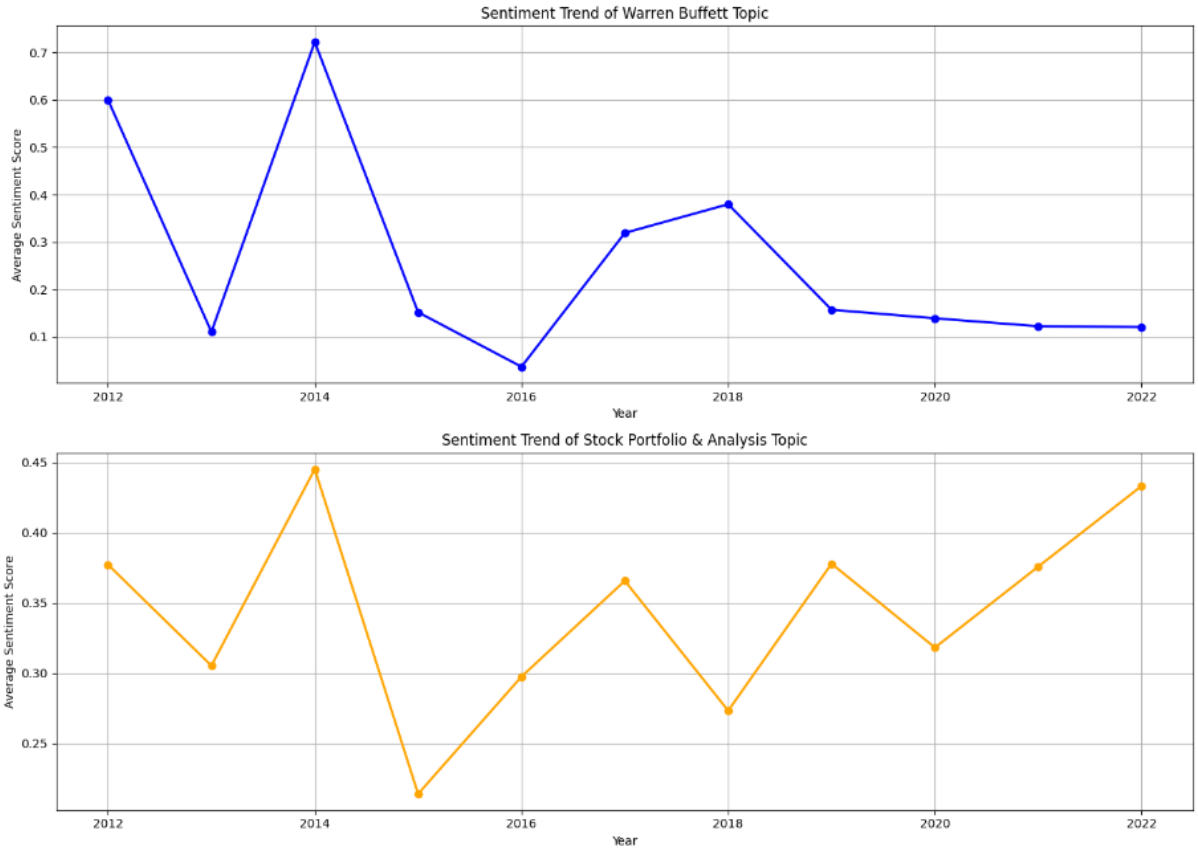
user_activity_topic2["z_score"] = stats.zscore(user_activity_topic2["post_count"])



- The Warren Buffett discussion has lower message counts and Z-scores, meaning users post at a relatively similar frequency, with no extreme outliers.
- The Stock Portfolio & Analysis discussion shows higher message counts and Z-scores, indicating a few users post at significantly higher rates, contributing disproportionately.

4.9 Sentiment Analysis of Topic Discussions

- Finally, I conducted sentiment analysis to assess the emotional tone of discussions within the two selected topics. I calculated average sentiment scores per year and plotted the trend over time.
- This analysis provided insights into whether discussions in the subreddit were becoming more positive or negative over time. Additionally, it allowed me to investigate whether sentiment shifts were influenced by major financial events, such as market crashes, economic policies, or significant investments by Warren Buffett.



- Warren Buffett discussions show sharp sentiment fluctuations, reacting strongly to specific events and investments.
- Stock Portfolio & Analysis discussions have a more stable and gradually increasing sentiment trend, reflecting broader market conditions.
- Warren Buffett discussions show a declining sentiment trend post-2020, suggesting increasing neutrality or criticism, while Stock Portfolio & Analysis sentiment steadily rises, reflecting growing optimism and confidence in investment strategies.

5. Predicting Future Super Users and Their Engagement Patterns

- Understanding super users helps in analyzing engagement trends, influence, and long-term participation. This section focuses on:
    1. Ranking existing super users based on their engagement.
    2. Predicting future super users using machine learning models.

  By analyzing past trends, I can pinpoint key contributors and forecast which users are likely to emerge as highly influential in the future. This helps in understanding evolving engagement patterns within the subreddit.

5.1 Ranking Super Users Based on Engagement Score

- An engagement score was computed to rank users based on activity, impact, and influence. The ranking was determined using:
  - **Post Count** – Total posts and comments made.
  - **Upvote Score** – Net upvotes and downvotes received.
  - **Reply Count** – Number of responses received.
  - **Unique Topics** – Number of different topics a user participated in.
  - **Average Sentiment** – Overall tone of the user's contributions.

Top 5 Super Users (Based on Engagement Score)

Author	Post Count	Upvote Score	Reply Count	Unique Topics	Avg Sentiment
Zurevu	2040	2224	2040	2	0.4086078431372549
[deleted]	1963	2055	1963	2	0.09814569536423841
StockTrendsBot	62	78	62	2	-0.6596774193548387
minorthreatmikey	26	73	26	2	0.14038461538461539
red5145	22	81	22	2	0.16909090909090907

- **Key Insights from the Ranking**
  - Users with a high upvote score and frequent replies tend to shape key discussions.
  - Those engaging in multiple topics act as discussion hubs, connecting different aspects of investment conversations.
  - The ranking helps identify the most influential contributors who actively drive subreddit engagement.

5.2 Predicting Future Super Users Using Machine Learning

- To forecast rising super users, I trained a machine learning model based on past engagement trends.
  - Defining Past Super Users - Users in the top 5% of post count before 2020 were labeled as super users.
  - Training a Random Forest Model - Features included Post Count, Upvote Score, Reply Count, Unique Topics, and Avg Sentiment.
  - Applying the Model to Recent Users (2020-2023) - The model identified users exhibiting similar engagement patterns as past super users.

Predicted Future Super Users (2020-2023)

Author	Post Count	Upvote Score	Reply Count	Unique Topics	Avg Sentiment
AutoModerator	18	18	18	1	0.23833333333333334
ElectrikDonuts	21	82	21	2	0.2628571428571429
Ibelieveyou2	19	17	19	2	0.41789473684210526
InvestmentClub-ModTeam	123	123	123	1	0.4741463414634146
Martenus	14	36	14	2	0.33428571428571424
Tombfz	11	21	11	1	0.09636363636363637
WallStResearch-Bot	30	30	30	1	0.341
destroyerodin	20	35	20	2	0.378
dopexile	12	17	12	1	-0.05916666666666667
krisolch	20	20	20	1	0.26549999999999996
kurt_FI	14	22	14	1	0.42214285714285715
minorthreatmikey	26	73	26	2	0.14038461538461539
red5145	22	81	22	2	0.16909090909090907
rifleman209	13	20	13	2	0.2623076923076923
shadowpawn	13	31	13	1	0.18692307692307694

- **Key Insights from Predictions**
  - Identifying rising contributors helps foster community engagement proactively.
  - Comparing past and predicted super users helps analyze shifting engagement trends.
  - This analysis allows communities to nurture participation and sustain subreddit activity over time.

6. Summary and Conclusion

This analysis of the InvestmentClub subreddit provided deep insights into user engagement, discussion trends, and community structures. By applying data cleaning and processing techniques, I ensured that the dataset was well-structured and free from inconsistencies, allowing for accurate analytical interpretations.

The graph visualizations offered a detailed understanding of user interactions and network dynamics. The directed interaction graph showcased how discussions were concentrated among a few highly engaged users. The centrality measures helped identify key influencers, highlighting users with high degrees of connectivity, bridging different discussion groups, and efficiently spreading information.

The super user analysis revealed patterns of engagement by identifying the most active contributors and their impact on discussions. Topic modelling allowed me to categorize conversations into meaningful groups, helping understand which investment topics attracted the most discussions. The community detection analysis further emphasized how discussions were structured, showing the formation of distinct user clusters.

Comparative analysis of two major topics, "Warren Buffett" and "Stock Portfolio & Analysis," provided insights into how discussions evolved over time. The Rich Club Coefficient highlighted whether high-degree users formed exclusive groups, while the sensitivity analysis examined how subreddit engagement was affected by the removal of key users.

To predict future super users, I leveraged machine learning models trained on past engagement data. This helped forecast which users were likely to emerge as key contributors in the community.

Key Takeaways

- User engagement is concentrated among a small number of highly active users who influence discussions.
- Stock Portfolio & Analysis discussions are more inclusive, while Warren Buffett discussions are dominated by a few key users.
- Community detection analysis showed that investment discussions tend to form distinct, well-connected clusters.
- Sentiment analysis indicated fluctuating emotions in Warren Buffett-related discussions, while Stock Portfolio & Analysis showed a steadily positive trend.
- Predicting future super users can help identify influential contributors early and sustain subreddit engagement.

This study demonstrates the power of network analysis, machine learning, and sentiment analysis in understanding online communities. By identifying key users, tracking discussion trends, and predicting engagement patterns, this research provides valuable insights into how investment communities evolve and how participation can be encouraged to foster meaningful discussions.