WallStreetPulse: Analyzing the GameStop (GME) Event Using Large Language Models

Abstract

The GameStop (GME) event in January 2021, catalyzed by the Reddit community r/WallStreetBets, underscored the significant influence of online investor coordination on financial markets. This highlights the imperative of comprehending the dynamics of online communities and their impact on financial phenomena. Our project employs large language models (LLMs) to examine the GameStop event, with a specific focus on identifying influential users, tracing information dissemination, and tracking the evolution of ideas within r/WallStreetBets.Through a synthesis of data collection, language modeling techniques, and network analysis, we aim to construct a comprehensive model illuminating the intricate social interactions and information cascades inherent in online platforms.

1. Introduction

The GameStop (GME) event of January 2021, orchestrated by retail investors from the r/WallStreetBets subreddit, exemplified the considerable sway of online investor communities over financial markets, challenging conventional stock market paradigms. Understanding and dissecting these events are imperative for stakeholders such as regulators, financial institutions, and investors alike. However, traditional analysis methodologies often struggle to capture the intricacies of online communities. In response, our project harnesses the power of Large Language Models (LLMs) to delve into the complexities of the GameStop phenomenon, scrutinizing influential users, information propagation, and the evolution of ideas within r/WallStreetBets. By integrating advanced data collection techniques, language modeling, and network analysis, we aim to transcend the limitations of traditional data analysis methods and provide a nuanced understanding of online social dynamics and information dissemination.

Motivation:

The GameStop event highlighted the potential influence of online communities and social media on financial markets, demonstrating the power of coordinated retail investor actions. Our goal is to leverage the capabilities of large language models to understand the underlying factors that contributed to this event, including the flow of information, user activities, and sentiment within the r/WallStreetBets community.

Challenges:

1) **Data Retrieval**: The Reddit API only provides access to the most recent 1,000 posts, limiting our ability to gather historical data from r/WallStreetBets during the critical time period of the GameStop event.

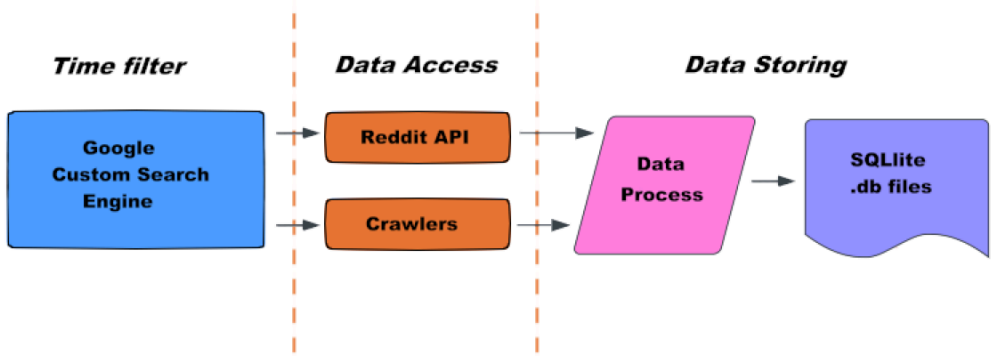
2) **Modeling as a Social Network**: The r/WallStreetBets subreddit can be viewed as a complex social network, with users interacting, sharing information, and influencing each other's decisions. Modeling the flow of information and the dynamics within this network is a challenging task.

3) **Analyzing User Activities with Large Language Models**: Effectively prompting and querying large language models to extract valuable insights from user activities and discussions within the r/WallStreetBets community requires careful consideration and experimentation.

2. Data Collection and Preprocessing

To analyze the GameStop (GME) event on r/WallStreetBets, we initiated a detailed data collection process. We faced challenges with historical Reddit data due to API constraints that limit access to the most recent 1,000 posts. To overcome this, we employed a custom search engine API and developed a web crawler, enabling us to gather a broader range of historical posts and comments.

In the preprocessing phase, we refined the data by filtering out irrelevant content through keyword searches and removing duplicates. This resulted in a clean dataset consisting of 19 posts and 74,661 comments, covering key discussions from January 1, 2021, to February 28, 2021. This dataset provides a comprehensive view of the critical conversations during the GameStop event.



3. Modeling Approaches We employed three modeling approaches:

3.1. User-to-User Model (Virus in Network Model)

The Virus in Network Model conceptualizes the spread of information within the r/WallStreetBets community as a viral contagion through a network during the GameStop event. This model employs a network graph where users are represented as nodes and interactions among them as edges. Each node, or user, has a probability of "infecting" its neighbors with information, simulating how information spreads through user interactions, primarily measured via comments.

Mechanics of the Model

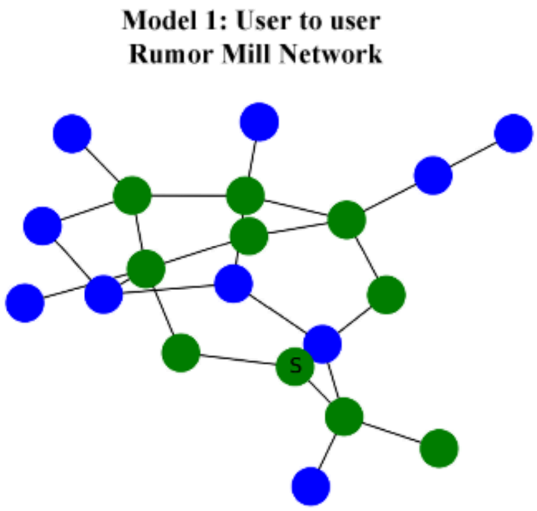
In the model, each node has a 50% probability of infecting adjacent nodes every eight-hour timestep, and a 5% probability of losing interest in the topic (i.e., becoming "immune" or resistant to further infection). This dynamic continues until a predefined number of iterations are reached, representing the simulation's end. The initial infection starts from "seed" users, identified as central influencers, and spreads based on these transmission probabilities.

Parameters and Verification

It’s important to highlight that the specific parameters of the model (such as the 8-hour timestep and the probabilities of infection and disinterest) are adjustable. These can be calibrated to more closely reflect the actual dynamics observed during the GameStop event, thereby enhancing the model's accuracy and predictive power. Rigorous verification against actual event data is essential to validate the model's outputs and ensure its relevance as an analytical tool.

Network Graph and Visualization

The network graph is constructed by parsing and analyzing user interaction data from the subreddit. We use Python scripts for sorting these interactions and simulating the spread of information. The final visualization represents the diffusion of information among users, highlighting how certain nodes (users) play critical roles in information propagation.
Note: The full visualization of the actual event is not displayed due to its complexity—with over 40,000 nodes it would be visually overwhelming. Instead, we provide an example visualization in Figure XXX, where green nodes represent "infected" users, blue nodes represent "uninfected" users, and 'S' marks the "seed" users.
Fig. XXX. *Graph Visualization of an example Virus in a Network Model, with green nodes as infected nodes, blue as uninfected, and “s” as a “seed” user.*



After collecting data from Reddit and formatting it to the model and processing the data, a sorted list of the most influential users was gained. The top 10 data points are presented in **Table XXX** below. These data reveal who are the most influential users during the event. Interestingly, the user Interaction Score dropped below 100 past the 18th user on the list. This implies that, while the event was caused by a large group of people convening on a single mindset, only a select few individuals were genuinely influential in the making of the event. By studying the actions of these users in future work, models could be trained on how to better influence groups of people through group think.

Table XXX. *Top 10 Influence Score results from Virus in Network Model simulation by user. Usernames returned as "none" excluded, removing one point with an Influence Score of 24648.*

Username	**Influence Score**
mcuban	74797
zjz	69061
does-it-matter	36573
dumbledoreRothIRA	18300
benafflecks	7106
TheHappyHawaiian	6961
The-Crazed-Crusader	5868
WorriedBanker	2287

Fragsworth	1555
thewaybaseballgo	1256

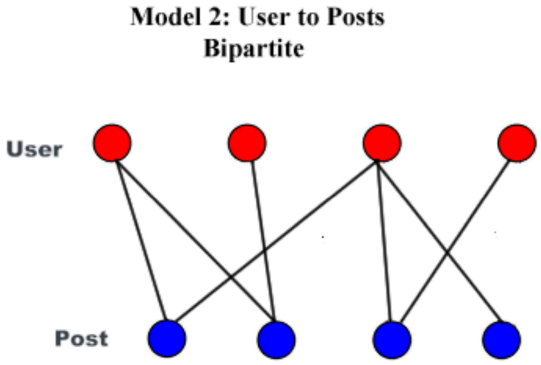
3.2. User-to-Posts Model (Bipartite Graph)

In the User-to-Posts model, we analyze the interactions between users and their respective posts through a bipartite graph. This model structure is particularly useful for visualizing and quantifying the extent of user engagement with specific content within r/WallStreetBets during the GameStop event. Here, the graph's two distinct sets of nodes represent users and posts, respectively, while the edges indicate the volume of engagement, primarily measured through comments.

Construction of the Bipartite Graph

The construction of the bipartite graph began by parsing interaction data collected from Reddit, specifically focusing on user engagement with posts. We mapped these interactions to quantify and illustrate the influence of users based on how their posts engaged the community. An example visualization of a Bipartite Graph is presented in **Fig. XXX.** below.

Fig. XXX.



Analysis Using Centrality Measures

To determine the most influential users, we employed centrality measures within the graph. These measures help in identifying users whose posts have not only garnered a lot of interactions but have also influenced the flow of discussion significantly within the community. By doing so, the model can identify individuals who might have played crucial roles in spreading key narratives during the moments of the GameStop event.

Results from the Bipartite Graph Model

Our analysis has highlighted several key users who were particularly influential in terms of engaging others with their posts. The top 10 users based on the engagement scores are listed below, providing insights into who drove the most interaction within r/WallStreetBets during the event:

Table 1: Top 10 Users by Engagement Score in the User-to-Posts Model

Rank	Username	Engagement Score
1	mcuban	1,208,814
2	TheHappyHawaiian	551,395
3	Nungie	528,398
4	The-Crazed-Crusader	415,545
5	Riverman786	321,924
6	audion00ba	276,927
7	B20Bravo	274,514
8	verascity	274,164
9	hypnoticfire69	242,112
10	probablyblocked	234,492

3.3. User in Communities Model (Community-Affiliation Graph Model for Overlapping Network Community Detection)

The User in Communities Model utilizes a community detection algorithm to identify overlapping sub-communities within the r/WallStreetBets subreddit. This model constructs a community-affiliation graph, where users are depicted as nodes and their interactions as edges, to explore the complex social structure of the subreddit.

Community Detection and Analysis
The model employs algorithms that maximize a quality metric such as modularity to iteratively detect communities. This approach allows us to see how users may belong to multiple communities, reflecting the multifaceted affiliations and roles they play within the subreddit. This method is particularly effective in revealing influential users within specific communities by analyzing their centrality and the density of their interactions.

Dynamic Model Adaptation
In our analysis, we use topic-based clustering to assign users to communities. Each post and comment is evaluated using OpenAI's GPT-3.5 API to determine its alignment with existing topics or the need for creating new community clusters. This process ensures that the community graph evolves dynamically as new content is added, maintaining the model's relevance and adaptability to the subreddit's changing dynamics.

Visualization and Insights
The resulting graph continuously updates to reflect the community's evolving interactions and discussions. Our visualization efforts focus on how these interconnected user nodes influence the subreddit's overall discussion landscape and dynamics. This model not only identifies but also highlights the roles and influence levels of users within various overlapping communities.

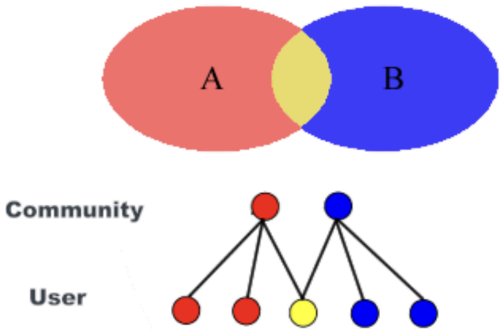


Figure 3: Community to Users Graph

4. Identifying Influential Users

In our quest to identify the most influential users within the r/WallStreetBets community during the GameStop event, we employed a comprehensive approach that synthesized insights from

three distinct models: User-to-User, User-to-Posts, and User in Communities. This methodological framework afforded us a nuanced understanding of influence dynamics and information dissemination patterns. To integrate the outcomes of these models effectively, we adopted a weighted summation technique, wherein scores were normalized to ensure consistency and prevent any single model from exerting undue influence.

We calculated influence scores as follows:

Influential User Score = $A \times normal\ (model1\_score) + B \times normal\ (model2\_score) + C \times normal\ (model3\_score)$

Here, *model1_score*, *model2_score*, and *model3_score* represent the scores generated by the User-to-User, User-to-Posts, and User in Communities models, respectively. The normalization function, normal(), scales these scores to a standardized range, typically between 0 and 1.

The coefficients A, B, and C allow for the customization of each model's contribution to the final influence score, thereby accommodating diverse analytical objectives. For instance, if prioritizing information dissemination, a higher weight might be assigned to the User-to-User Model (A > B, C).

Subsequently, users were ranked based on their final influence scores, enabling the identification of the most prominent influencers within the r/WallStreetBets community.

5. Results and discussion

Our study identified a list of influential users within the r/WallStreetBets community during the GameStop event, with corresponding scores indicating their contribution to information spread. This list provides an initial quantification of influence based on interactions and engagements within the community, as modeled through our three distinct approaches: User-to-User, User-to-Posts, and User in Communities.

However, our result is under the process of further verification. One potential validation approach is comparing our identified influencers with publicly recognized influential figures from the event, although limited data online makes this challenging. Additionally, analyzing user earnings during the event using LLMs could offer insights into the correlation between online influence and financial outcomes. In addition, applying complex network analysis and comparing our results with similar studies could provide deeper insights into the validity of our methods and findings.

6. Future works

For future research, extending the analysis beyond February 2021 to include data up to October 2021 could reveal the long-term effects of the GameStop event and its correlation with subsequent stock price movements. This extension would also allow us to explore the event's evolution over time.

We also plan to explore additional ways to gain deeper insights into the GameStop event discussions on r/WallStreetBets. Specifically, we intend to:

1. **Enhance Post Analysis:** Develop methods involving Large Language Model analysis to assess the strength of calls to action within each post.
2. **Apply PageRank Algorithm:** Use the PageRank algorithm to visualize better the influence of individual users within the subreddit's network, giving the rank not only based on direct interactions but also on their centrality within the broader information flow.

Additionally, employing specialized language models like FinGPT, which are fine-tuned on financial data, might enhance our ability to extract more precise insights from the discussions on r/WallStreetBets. Experimenting with different prompting strategies could also refine the granularity of our analysis, potentially revealing more nuanced patterns of influence and information dissemination.

Further development of data visualization techniques would assist in better illustrating the complex structures of online community networks and the flow of information, making our findings more accessible and understandable to a broader audience. Applying our methodologies to other online communities across different domains could also validate the generalizability and effectiveness of our approach in various contexts.

References

- Dong, Zihan, et al. 2024, *FNSPID: A Comprehensive Financial News Dataset in Time Series*, https://arxiv.org/abs/2402.06698. Accessed 20 Mar. 2024.
- J. Yang and J. Leskovec, "Community-Affiliation Graph Model for Overlapping Network Community Detection," 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 2012, pp. 1170-1175, doi: 10.1109/ICDM.2012.139. keywords: {Communities;Image edge detection;Reliability;Organizing;Collaboration;YouTube;Community detection;Overlapping communities},

- Stonedahl, F. and Wilensky, U. (2008). NetLogo Virus on a Network model. http://ccl.northwestern.edu/netlogo/models/VirusonaNetwork. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.
- Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.
- Xia, Mengzhou, et al. 2024, *LESS: Selecting Influential Data for Targeted Instruction Tuning*, https://arxiv.org/abs/2402.04333. Accessed 20 Mar. 2024.