Week 5: Qualitative Analysis of Wallstreet Bets

Nate Bachmeier

TIM-7211: Research Design Methodologies

January 24, 2021

Northcentral University

# Qualitative Analysis of Wallstreet Bets

Wallstreet Bets (WSB) is a self-described community of “2.1 million degenerates” that is “like 4chan found a Bloomberg Terminal (WSBMod, 2021).” Members of this subreddit share their investment strategies via memes that integrate their unique vocabulary (see Table 1). There is value in understanding this group as they represent the retail investor voice. Their large group also deploys highly leveraged strategies by purchasing hundreds of inexpensive call options. For instance, they have recently banded together to push GameStop from $25 to $78 (Vena, 2020).

Table 1: Example Phrases

|  |  |  |
| --- | --- | --- |
| Phase | Translation | Use-Case |
| OP just YOLO’d GME | The original poster put their life savings into Gamestop | Describes an ultra-high risk bet |
| This is the way | I approve | Positive response to a comment |
| Stonks only go up | You are mistaken | A negative response to a comment |
| The autism is strong with this one | Promote this user to a moderator | A positive or negative response |
| Loss Porn | Sharing evidence of a failed YOLO | A member has lost all of their money |

# Problem Statement

A formal analysis of the community does not exist, and this introduces challenges in understanding internal structures. For example, Hertz Rentals was entering into bankruptcy in June 2020, until WSB began promoting its stock, raising its value 825%, and enabling a path forward (Fox, 2020). This incident raises numerous questions around identifying key players and related event timelines. Along with the Gamestop and Hertz examples, how many other stories exist in this repository? What starts the snowball that becomes these avalanches? Answering these questions requires dissecting their language and measuring the evolution over time.

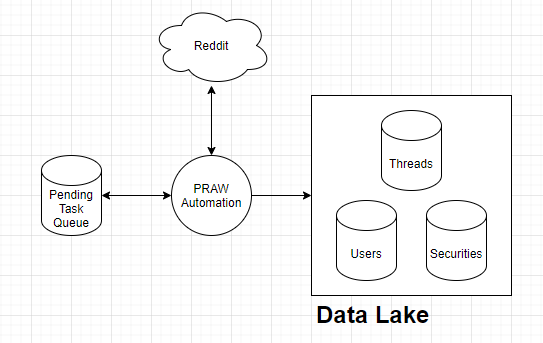
# Purpose Statement

Examining the evolution of the subreddit’s internal structures will provide value both academically and financially. Today, the community shows signs of a strong group-mentality, but who is leading that group? After identifying those leaders, there needs to be an analysis of their particular characteristics. For instance, are those accounts controlled by individual users or businesses that operate ‘pump and dump’ scams. Moderators actively remove malicious content that blatantly violates their rules, but more sophisticated attackers might exist. Finally, an exploration of discussed investment instruments (e.g., stocks and bonds) will help build the timelines.

# Data Collection Plan

An automation process needs to exist that downloads the WSB channel, its threads, and its users’ metadata (see Figure 1). This action requires communicating with Reddit’s RESTful (Representational State Transfer) web services. Numerous open-source libraries like Python’s Reddit API Wrapper (PRAW) already interact with these endpoints.

Figure 1: Data Collection Architecture



The collection process will need to collect an enormous volume of data, requiring days to download. Since each thread can be retrieved independently, decomposing the data transfer into independent tasks allows parallel processing. A service like Amazon Simple Queuing Service (SQS) or Apache ActiveMQ can maintain the state and distribute the work across multiple compute instances. After downloading the semi-structured data (see Figure 2), a data lake will persist the data for further analysis.

Table 2: Important Data Properties

|  |  |
| --- | --- |
| Property Name | Description |
| Id | Unique identifier for the comment |
| Thread Id | Unique identifier of the thread |
| Author | The content creator’s alias |
| Body | The content in Markdown format |
| Score | Upvotes minus Downvotes |
| Created UTC | Initial time |
| Replies | References to the responses |
| Edited | Flag to denote the body changed |
| Distinguished | Flag to denote the comment is exceptional |

# Data Analysis Plan

Quantitative methods seek to measure a phenomenon’s impact and generalize those insights to similar scenarios (Creswell, 2014). However, this research topic focuses on a specific community and is not generalizable. DeCarlo (2018) claims that social studies often encounter this limitation and turn to qualitative techniques. There is a frequent misconception that qualitative methods are inferior to their quantitative counterparts. However, academic research shows it is equally effective when used correctly (Sarma, 2015).

## Message Selection Process

First, a filtration process will remove any negative score messages, as the WSB users reject those opinions. Next, parsing the body property produces unformatted phrases, securities, emojis, and its sentiment. User’s comments can discuss any subject, necessitating another filter to remove general chit-chat. After that, a Natural Language Processing (NLP) algorithm seeks to cluster the phrases into a handful of buckets. Finally, the buckets assign comment annotations to normalize the language.

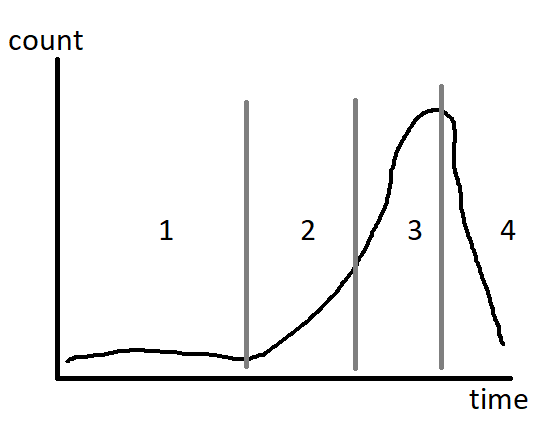
## NLP Algorithm Design

The Natural Language Tool Kit (NLTK) is an open-source library for advanced text analysis. Boorugu & Ramesh (2019) demonstrates a similar technique by first encoding the message into a bag of words. Next, they filter the bag using a Term Frequency filter to scale an individual term's importance. For instance, words like “the” should have less influence than a keyword like “GameStop.” Afterward, Gensim computes a similarity score between the documents. Finally, this score becomes the distance argument into a standard k-Nearest Neighbors (KNN) process.

## Assessing Interest and Sentiment

Afterward, counting the frequency of security instruments mentioned over time forms the high-level timeline. For instance, Hertz declared bankruptcy on May 22 and canceled this decision on June 5, 2020. During this period, both aggregate sentiment and mentions must have fluctuated. Those inflection points require a group of essential users that drive the momentum and encourage their peers to follow. Identifying these users involves assessing their comment creation time, sentiment, and various account metadata.

Figure 2: Example Inflection Points in Mentions



## Comparing Leaders

Finally, an analysis of the leader’s comment annotations evaluates the similarity between them. If there is a high degree of similarities, then this could be evidence of duplicate accounts. Alternatively, there could be an echo chamber, which causes the same similar wording across legitimate accounts. For instance, catch-phrases such as “this is the way” are frequently repeated. This risk introduces the need to have different annotation weights while comparing account similarities.

# References

Boorugu, R., & Ramesh, G. (2020). A survey on NLP based text summarization for summarizing product reviews. *International Conference on Inventive Research in Computing Applications*, (pp. 352-356). Coimbatore, India. doi:10.1109/ICIRCA48905.2020

Creswell, J. (2014). *Research design: Qualitative, quantitative, and mixed methods approach.* Thousand Oaks, CA: Sage Publishing, Inc.

DeCarlo, M. (2018). *Principles of Sociological Inquiry: Qualitative and Quantitative Methods.* Blackstone Textbooks. Retrieved from https://scientificinquiryinsocialwork.pressbooks.com/chapter/12-1-experimental-design-what-is-it-and-when-should-it-be-used/

Fox, M. (2020, June 5). *Hertz skyrockets since filing for bankruptcy as Robinhood traders pile into HTZ*. Retrieved from Business Insider: https://markets.businessinsider.com/news/stocks/hertz-stock-price-skyrockets-since-filing-bankruptcy-traders-global-holdings-2020-6-1029285231

Sarma, S. (2015). Qualitative research: examining the misconceptions. *South Asian Journal of Management, 22*(3), 176-191.

Vena, D. (2020, January 23). *WallStreetBets Declares Victory as GameStop Stock Soars Over 50%*. Retrieved from MSN Money: https://www.msn.com/en-us/money/markets/wallstreetbets-declares-victory-as-gamestop-stock-soars-over-50percent/

WSBMod. (2021, January 24). *Wallstreetbets*. Retrieved from Reddit: https://www.reddit.com/r/wallstreetbets/