Journal of Open Humanities Data

Reddit Dataset on Meme Stock: GameStop

DATA PAPER

JING J. HAN

ubiquity press

ABSTRACT

This dataset includes one-year Reddit posts, post metadata, post sentiments, 57 post meta-features and post comments threads from several subreddits around a meme stock: GameStop. These subreddits are dedicated to the discussion of GameStop stock and the social movement of global wealth transfer that the event GameStop stock short squeeze initiated in January 2021. The subreddits included are r/GME, r/superstonk, r/DDintoGME, and r/GME Jungle. The whole dataset is stored in eight CSV files and four HTML files.

CORRESPONDING AUTHOR:

Jing J. Han

Klein College of Media and Communication, Temple University, Philadelphia, US

jing.han@temple.edu

KEYWORDS:

GameStop; Reddit; online social movement; meme

TO CITE THIS ARTICLE:

Han, J. J. (2022). Reddit Dataset on Meme Stock: GameStop. *Journal of Open Humanities Data*, 8: 20, pp. 1–5. DOI: https://doi.org/10.5334/johd.85

(1) OVERVIEW

The collection of this dataset was inspired by the short squeeze event on GameStop stock initiated by retail investors in January 2021. A short squeeze is an unusual condition that triggers rapidly rising prices in a stock or other tradeable financial instruments. For a short squeeze to occur, the financial instrument must have an unusual degree of short sellers holding positions in it. The short squeeze is triggered when short sellers coincidentally cut losses and exit their positions (Mitchell, 2021). At its height, the pre-market value for GameStop stock was more than $500 per share (Wikipedia, 2022). GameStop stock is a meme stock that refers to the shares a company has gained online following through social media platforms. These online communities can build hype around a meme stock through narratives and conversations, which reflect public opinions of the stock (Hayes, 2022).

Reddit has been the primary platform retail investors use to communicate with each other, including sharing and discussing news from social media and mainstream media, personal trading histories, memes, technical analyses, and strategies to facilitate global wealth transfer. The goal of the Reddit community online movement was colloquially named "Mother of All Short Squeezes" (MOASS) (Anand & Pathak, 2021; Betzer & Harries, 2022). MOASS exemplifies the populist intent of an online social movement observed in the dataset. The realization of MOASS's goals requires participation from every community member. However, differing opinions about how their goal should be achieved and what kind of community culture should be constructed have split the community into four subreddits. Specifically, the creation of r/superstonk was born of frustration about r/wallstreetbets, which received mainstream media attention at the beginning of the short squeeze. The subreddit r/superstonk was created, driven by the lack of focus on achieving the common goal and the concern on the intention and conduct of moderators on r/wallstreetbets. Its profile banner, "Power to the Shareholders" distils its populist belief in achieving the common goal. However, the integration with meme culture on r/superstonk has distanced community members who are motivated to achieve the common goal with a more serious and less memetic attitude. The community migration into r/GME, r/DDintoGME, r/GMEJungle was the result of this cultural disagreement. Furthermore, community migration does not follow a linear progression nor does it suggest that the community is conflicted and divisive. Instead, it reflects the influence of Reddit's features on the organization of the community: individuals' content curation on Reddit is structured by topics. Users on Reddit curate their content by following different subreddit communities. Thus, this dataset will help study online social movements and its relationship with online culture.

The collection of data was motivated by the continuous actions of community members pursuing the realization of MOASS. During the data collection period, several changes in communication patterns and communication tactics occurred, driven by both internal and external events, such as community disagreement on ways of realizing common MOASS goals, and episodic mainstream media attention.

The dataset on r/superstonk has 560,125 posts with an average word count of 15 and a standard deviation of 13 rounded to the nearest integer. The dataset on r/GME has 1,033,236 posts with an average word count of 14 and a standard deviation of 13 rounded to the nearest integer. The dataset on r/GMEJungle has 39,634 posts with an average word count of 15 and a standard deviation of 12 rounded to the nearest integer. The dataset on r/DDintoGME has 5,498 posts with an average word count of 16 and a standard deviation of 13 rounded to the nearest integer. The four HTML files on explorative data analyses demonstrate the first 12 variables (id, title, url, score, author, number of comments, date, flair, negative sentiment, positive sentiment, neutral sentiment, and compound sentiment), their interactions, and correlations from the dataset files ending with "features."

REPOSITORY LOCATION

Context

This dataset was produced as part of an ongoing research project[1] that studies the communication patterns of subreddit communities around meme stocks and their belief in using meme stocks to facilitate a global wealth transfer movement. It has not been used in any publication yet.

1 Coding notebooks from this project will be shared publicly in the future.

(2) METHOD

The post ID, title, URL, score, author, number of comments, date, and flair (community-defined content filter) were collected by using Pushshift Reddit API (Baumgartner, 2018). The post comments were collected by using the Python Reddit API Wrapper, PRAW (Boe, 2021). Each post's sentiment scores were calculated using VADER (Hutto & Gilbert, 2014) with a customized dictionary that reflects the common emojis used in these subreddits. 57 meta-features on post titles were produced by using the spaCy large English model (Honnibal et al., 2020). The explorative data analyses are generated by pandas profiling (Brugman, 2019) and sweetviz (Bertrand, 2022).

STEPS

I used pushshift to collect post titles and post metadata. Next, I used PRAW to collect post comments. The customized VADER dictionary assigned the "gem stone", "gorilla", different skin tones of "raising hands", "rocket", different versions of "moon", and different skin tones of "open hands" emojis to score four, which is the highest score in VADER, signifying high positive sentiment. The emoji "crayon" was assigned a score of one, reflecting a moderately positive sentiment. The distinctive emoji uses reflect the communication and language patterns in these subreddits. For example, the "gem stone" emoji means "diamond hands", which describes an investor who refrains from selling an investment despite downturns or losses. The combination of "rocket" emoji and "moon" emoji means "going to the moon", which describes when the price of a financial instrument is rising off the charts.

QUALITY CONTROL

The values collected from Pushshift, such as scores and number of comments, only reflect the values when the data was collected. There might be a discrepancy between the values collected and the real-time values. The customized update on VADER dictionary only includes commonly agreed-on emoji used by the GameStop retail investors. These particular emoji uses are also shared by the larger communities associated with the mentality of the meme culture. The pre-processing results on post titles are included in 57 meta-features, which are viable for future analyses, such as creating further features.

(3) DATASET DESCRIPTION

OBJECT NAME

Reddit Dataset on Meme Stock: GameStop

https://doi.org/10.7910/DVN/TUMIPC

FORMAT NAMES AND VERSIONS

CSV; HTML; Version 2.0

CREATION DATES

2022–02–15 — 2022–04–26

DATASET CREATORS

Jing Han

LANGUAGE

English

LICENSE

REPOSITORY NAME

Dataverse

PUBLICATION DATE

2022–07–09

(4) REUSE POTENTIAL

The 74 variables in this dataset provide opportunities for future analyses, such as creating further features during exploratory analysis and future studies. For example, the variable post flairs can be used as post labels for text classification research. Researchers who are interested in understanding online communication patterns could use this labeled dataset to train a classifier and apply multiclass or multilabel inference on the comment threads. The results of text classification research could also be used to understand the communication processes of these subreddits. The relationship between communication processes and the effects of the online social movement (MOASS) could be studied by performing a time series analysis on the dataset and analyzing mainstream media's attention on the movement. Furthermore, word count, stop word count, word count after cleaning, and speech tagging would be useful for named-entity recognition and online language studies. The results of studying online language use contained in the dataset would be helpful understanding the community culture of these subreddits, which could contribute to the studies on meme culture and broadly, online culture. Using public sentiment to harness the power of public opinion, research has outlined methods for analyzing commercial interests. For example, researchers have studied the relationship between public sentiment on social media platforms and market impact (Nguyen & Shirai, 2015; Audrino et al., 2020). S & P Dow Jones Indices includes a social media sentiment factor (S&P Global, n.d.). Sentiment annotation on this Reddit dataset using VADER with a customized dictionary could provide a baseline comparison for researchers interested in using sentiment as a variable to study the processes and effects of public sentiment. Specifically, the sentiment annotation could assist studies on the relationship between public sentiment and price fluctuations of stock, between public sentiment and public opinion.

The Reddit dataset generated during the GameStop short squeeze stands out from other Reddit corpus because of its socio-economic relevance. The social movement following the event demonstrates the power of people and the long-term economic impact their actions had. Additionally, Reddit allows access to data via its API terms of use, which is more generously than other social media platforms (Reddit, 2016). Reddit's data structure and limited restrictions on posting content provide opportunities to study online language use, communication processes, public opinions, online culture, online communities, and online social movements.

ACKNOWLEDGEMENTS

I gratefully acknowledge Dr. Ryan Omizo's guidance and encouragement in creating this dataset. Publication of this article was funded in part by the Temple University Libraries Open Access Publishing Fund.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jing Han is responsible for conceptualization, data curation, methodology, and writing.

AUTHOR AFFILIATION

Jing J. Han orcid.org/0000-0003-3251-6549
Klein College of Media and Communication, Temple University, Philadelphia, US

REFERENCES

Anand, A., & **Pathak, J.** (2021). WallStreetBets against wall street: The role of reddit in the gamestop short squeeze. *IIM Bangalore Research Paper, 644*. https://repository.iimb.ac.in/handle/2074/20101. DOI: https://doi.org/10.2139/ssrn.3873099

Audrino, F., Sigrist, F., & **Ballinari, D.** (2020). The impact of sentiment and attention measures on stock market volatility. *Capital Markets: Asset Pricing & Valuation eJournal*. DOI: https://doi.org/10.2139/ssrn.3188941

Baumgartner, J. M. (2018). *Pushshift API*. Retrieved from https://github.com/pushshift/api (last accessed: 9 May, 2022).

Betzer, A., & **Harries, J. P.** (2022). How online discussion board activity affects stock trading: the case of GameStop. *Financial markets and portfolio management*, 1–30. DOI: https://doi.org/10.1007/s11408-022-00407-w

Bertrand, F. (2022). *Sweetviz*. Retrieved from https://pypi.org/project/sweetviz/ (last accessed: 22 July, 2022).

Brugman, S. (2019). *Pandas-profiling: exploratory data analysis for python*. Retrieved from https://github.com/pandas-profiling/pandas-profiling (last accessed: 22 July, 2022).

Boe, B. (2021). *PRAW: The python reddit api wrapper*. Retrieved from https://praw.readthedocs.io/en/v7.5.0/ (last accessed: 9 May, 2022).

Hayes, A. (2022). *Meme Stock*. Retrieved from https://www.investopedia.com/meme-stock-5206762 (last accessed: 9 May, 2022).

Hutto, C., & **Gilbert, E.** (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, *8*(1), 216–225). URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al. (2020). spaCy: Industrial-strength natural language processing in python. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.1212303

Mitchell, C. (2021). *Short squeeze*. Retrieved from https://www.investopedia.com/terms/s/shortsqueeze.asp (last accessed: 26 April, 2021).

Nguyen, T. H., & **Shirai, K.** (2015). Topic modelling based sentiment analysis on social media for stock market prediction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1*, 1354–1364. DOI: https://doi.org/10.3115/v1/P15-1131

Reddit. (2016). *API Terms*. Retrieved from https://www.reddit.com/wiki/api-terms (last accessed: 22 July, 2022).

S&PGlobal. (n.d.). *Social Media Sentiment – Indices*. Retrieved from https://www.spglobal.com/spdji/en/index-family/strategy/factors/social-media-sentiment/#overview (last accessed: 22 July, 2022).

Wikipedia. (2022). *GameStop short squeeze*. Retrieved from https://en.wikipedia.org/wiki/GameStop_short_squeeze (last accessed: 22 July, 2022).

TO CITE THIS ARTICLE:
Han, J. J. (2022). Reddit Dataset on Meme Stock: GameStop. *Journal of Open Humanities Data*, 8: 20, pp. 1–5. DOI: https://doi.org/10.5334/johd.85

Published: 24 August 2022