This GameStop Reddit Dataset readme.txt file was generated on 08/31/2023 by Jing Han

---
GENERAL INFORMATION
---

1. Title of Dataset: GameStop Reddit Dataset

2. Author Information:

Principal Investigator Contact Information
Name: Jing Han
Institution: Temple University
Address:
Email: jing.han@temple.edu
ORCID:

Associate or Co-investigator Contact Information
Name: Andrew Iliadis
Institution: Temple University
Address:
Email: andrew.iliadis@temple.edu
ORCID:

3. Date of data collection (single date, range, approximate date): <suggested format YYYYMMDD>
r/DDintoGME: 2021/04/07 – 2021/12/31
r/GMEJungle: 2021/07/14 – 2021/12-31
r/superstonk: 2021/04/15 – 2021/12-31
r/GME: 2021/01/04 – 2021/12/31
The difference in start dates is from different subreddit inception time and Reddit data availability.

4. Geographic location of data collection: On Reddit

5. Information about funding sources or sponsorship that supported the collection of the data: Not funded

---
SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data, or limitations of reuse:

Public Domain

DATA & FILE OVERVIEW
---

1. File list (filenames, directory structure (for zipped files) and brief description of all data files, add additional entries as necessary):

A. Filename: r/superstonk
Short description: This dataset contains metadata (id, url, author, number of comments, date, flair, post title) and sentiment scores of post titles in r/superstonk. Post titles have been processed into lower-case.

B. Filename: r/GME
Short description: This dataset contains metadata (id, url, author, number of comments, date, flair, post title) and sentiment scores of post titles in r/GME. Post titles have been processed into lower-case.

C. Filename: r/DDintoGME
Short description: This dataset contains metadata (id, url, author, number of comments, date, flair, post title) and sentiment scores of post titles in r/DDintoGME. Post titles have been processed into lower-case.

D. Filename: r/GMEJungle
Short description: This dataset contains metadata (id, url, author, number of comments, date, flair, post title) and sentiment scores of post titles in r/GMEJungle. Post titles have been processed into lower-case.

2. Relationship between files, if important for context:

These subreddits are thematically related and they can be considered as the offshoot subreddits of the subreddit r/wallstreetbets, which was at the center of the GameStop short squeeze event occurred at the beginning of 2021.

METHODOLOGICAL INFORMATION
---

1. Description of methods used for collection/generation of data:

Pushshift Multithread API wrapper was used to collect all subreddit data in the date range. This wrapper enables programmatically collecting Reddit data with Python scripts.

2. Methods for processing the data:

VADER sentiment analysis was used to generate sentiment scores. More information about VADER sentiment analysis can be found at this github repository. The post titles were processed to be lower-case and multiple spaces in post titles were substituted with a single space,

3. Describe any quality-assurance procedures performed on the data:

Jing Han compared a small sample of collected data with the corresponding data on Reddit and did not have any discrepancies. However, the datasets might contain data that are no longer available on Reddit.

4. People involved with sample collection, processing, analysis and/or submission:

Jing Han collected, processed, analyzed, and submitted these datasets.

DATA-SPECIFIC INFORMATION

Description of the variables:
1,687,255 Reddit posts across 4 subreddits were collected.
Each dataset has the same variables.
id is the Reddit assigned ID for each post.
url is the link of a post.
score is the number of scores (which is a combination of upvotes and downvotes) of a post.
author is the name of a post author.
num-comments is the number of comments a post received.
date is when a post was being submitted.
flair is the subreddit-specific filters to categorize posts.
processed_title contains post titles processed to be lower case. Multiple spaces in post titles were substituted with a single space.
neg is the amount of negative sentiment in a post title using VADER sentiment analysis.
pos is the amount of positive sentiment.
neu is the amount of neutral sentiment.
compound is the overall sentiment.