FINM 35000 Problem Set 4: Retail Investor Sentiment

Instructor: Joanna Harris, TA: Lisheng Su

Due November 18, 2022, 6PM

Submission instructions: please submit one copy of the assignment per group. The submission should include a PDF or Word document with the written results, tables and figures, a separate code file (in a programming language of your choosing) and all input files necessary for the code to run. Please make sure to write the names of all group members at the top of the writeup.

Supplementary Files: the data files and papers that you will need for this assignment can be found at https://www.dropbox.com/sh/hqpgbp2gm130mqw/AAD1fir3uOS4zMEkddchx_YDa?dl=0

1 Measuring Retail Investor Sentiment from WallStreetBets

1. Load the file `reddit_wsb.csv` into the programming language of your choice. This file contains posts from the subreddit WallStreetBets from September 29, 2020 to August 16, 2021. [1] Create a new column that stitches together the "title" and "body" columns (using only the title when the body is missing).

   (a) Report summary statistics for the number of words of the title+body column (going forward I will refer to the title+body as the post).

   (b) Plot the number of posts by month.

2. Load the dataset `crsp_daily.csv` and drop firms whose tickers are "YOLO", "BUY", "DD", "GO", and "GAIN".[2] Using the list of tickers from the CRSP data, record which tickers are mentioned in each post and report the number of posts that contain a ticker. Note that tickers are often preceded by a dollar sign (e.g. $GME).

3. Regress returns on an indicator for whether the ticker was mentioned in any WSB posts on a given day or on the prior day. Specifically, you will run two regressions:

$$R_{i,t} = \alpha + \beta \mathbb{1}\{Mention_{i,t}\} + \varepsilon_{i,t}$$

and

$$R_{i,t+1} = \alpha + \beta \mathbb{1}\{Mention_{i,t}\} + \varepsilon_{i,t}$$

[1] This data comes from https://www.kaggle.com/datasets/gpreda/reddit-wallstreetsbets-posts. The posts for which the "body" column is missing appear to be those where the body of the post is a photo, not text. For these posts, you can just focus on the title.

[2] This is the cleaning step used by Hu et al. (2021) to avoid mislabeling posts that contain these terms as being related to firms with these tickers.

4. Now restrict the returns data to only the stocks that are ever mentioned in the Reddit data. Run the following two regressions:

   $$R_{i,t} = \alpha + \beta Mentions_{i,t} + \varepsilon_{i,t}$$

   and

   $$R_{i,t+1} = \alpha + \beta Mentions_{i,t} + \varepsilon_{i,t},$$

   where $Mentions_{i,t}$ is the number of times stock i was mentioned on day t.

5. For each post, count the instances of the phrases, "buy," "sell," "hold" and "short." Construct the following measure of whether the post is positive or negative on the stock:

   $$RedditRec = \begin{cases} \frac{\#Buy+\#Hold-\#Sell-\#Short}{\#Buy+\#Hold+\#Sell+\#Short} & \text{if } \#Buy + \#Hold + \#Sell + \#Short > 0 \\ 0 & \text{otherwise} \end{cases}$$

   Run the following regression:

   $$R_{i,t+1} = \alpha + \beta RedditRec_{i,t} + \varepsilon_{i,t},$$

6. Comment on your results from parts 3-5.

7. Describe the shortcomings of this simple measure of sentiment. Describe a methodology you would use to better measure sentiment from the text of WallStreetBets posts.

2 Reading Response (Optional)

Note: This section is optional and can be submitted for bonus points.

Read Barber et al. (2021) and answer the following questions.

1. What is attention-induced trading?

2. What do you like about the paper? What do you think it could do differently?

References

Barber, B. M., Huang, X., Odean, T., and Schwarz, C. (2021). Attention induced trading and returns: Evidence from robinhood users. *Journal of Finance, forthcoming.*

Hu, D., Jones, C. M., Zhang, V., and Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Available at SSRN 3807655.*