Eric Kearney

CS-542 Final project proposal

November 8, 2022

Who knows best: r/wallstreetbets, r/stocks, r/investing?

I found a dataset on Kaggle of scrapped comments from investing related subreddits on reddit.com (credit to user HADI) stretching past for over five years. Two of them, r/investing and r/stocks, view themselves as more ’serious’ investing discussion boards, while the final subreddit, r/wallstreetbets, is the collection of people behind the now infamous Gamestop stock insanity.

My project idea is to identify a portfolio of ∼10 stocks, filter the comments from each subreddit based on these stocks, and perform sentiment analysis over time of these comments. Because the data are unlabeled, (and I don’t anticipate having anywhere near the amount of time or patience requried to go through and manually label thousands of reddit comments), I’ll treat this as an unsupervised problem, and use K-means clustering. I’ll then compare the subreddit’s sentiment with the actual performance of these stocks in order to try and determine which subreddit is the ’wisest’ (or perhaps rather, which one is the lest un-wise).

A couple decisions I have yet to make are whether each of the subreddits should have to share the same portfolio, or whether each should get their own tailored one. r/wallstreetbets will naturally be more inclined to discuss ’meme’ stocks than the other two subreddits, so while a shared portfolio may be easier for me, allowing each subreddit to have their own portfolio of the stocks they discuss the most will likely yield more interesting results.

If I have time, I’d also like to try using VADER, a tool I found online specifically tailored for Twitter sentiment alaysis and compare its results to my K-means clustering approach.