CIS545 Spring 2021 Final Project Proposal

Group Member: Bowen Tan, Xuanyang Wang

Data Source:

We are looking into Reddit WallStreetBets Posts as our data source. We found two complete datasets containing 1117296 and 42533 posts from kaggle separately. Since those 2 datasets share some common features such as score and title, but also have different features such as author in the first one and comment count in the second, we decided to use both of them.

Urls:
https://www.kaggle.com/unanimad/reddit-rwallstreetbets (First)
https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts (Second)

Project Plan:

Explain what you intend to study with your project.

We are going to find the relationships between the author/keywords in the title and the score/comments of the posts.

We would also try to identify the stock / fund that are mostly discussed within different periods of time based on the title.

Additionally, we could also look into the idiom of the users by analyzing the word frequencies of different users.

What is the ultimate objective?

Our ultimate objective is to find out the correlation between the user posts and score/comments, and find out what most users are interested in.

What types of models are you considering?

Since there are many categorical data in the dataset, we might try models such as adaboost and random forest.

Why is this project interesting?

As the WallStreetBets on reddit actually initialized the AME & GEM event which had enormous impact on the market weeks ago, it is interesting for us to see what the users in this particular part of the forum usually discuss, and what topics they normally focus on. We also think we can get some information about stocks and funds that are with great potential based on the score and comment number of the specific posts, and potentially make some profit from it :)

What challenges and obstacles might you anticipate with this project?

1. It is hard to distinguish keywords, especially the financial product names from the messy words in the title

2. Some of the features are highly unbalanced such as score and message body

3. The two datasets might have some inconsistencies.

4. We might be missing some relative information since we do not have detailed content of the comments.