

1. Choice of dataset

We will be using the [Reddit API](#) from where we will analyze the sentiment on the r/Wallstreetbets subreddit. For stock price information we are going to use the [TDAmeritrade API](#). By combining insights from the data we extract, we want to predict the bullishness or bearishness of the NYSE(New York Stock Exchange) and of particular stocks on the exchange.

2. Methodology

- a. **Data Preprocessing:** The Reddit API gives us access to post content and comments across the r/Wallstreetbets subreddit. By scraping the comments under each post, we can then pass them to our model without much preprocessing needed.
- b. **Machine Learning Model:** Determining the general health of the stock market i.e whether the market is a bullish or bearish market based on analyzing the general sentiments of the Redditors on Wallstreetbets. We'll be using Natural Language Processing for sentiment analysis
- c. **Evaluation Metric:** We are unable to use accuracy as an evaluation metric since we do not know the number of correct queries. We will decide on what to use as we learn about other metrics. Some metrics such as cross-validation accuracy might be the answer

3. Application

The user can input specific stocks they would like to inquire about and gauge how sentiments about that stock affected its health.