Final Project Proposal

1. Name of the project and team members

Project Title: Do Social Media Sentiments Drive U.S. Stock Market Reactions?

Team Members:

Mengyue Niu USC ID: 5926-7082-41
Yichen Yang USC ID: 4403-8375-34

2. What problem are you trying to solve?

This project investigates whether investor sentiments expressed on social media platforms influence short-term U.S. stock price movements and trading activity. The goal is to quantify how collective emotions in online investor communities translate into observable market reactions, providing data-driven insight for short-term investment decisions.

Problems we are trying to solve:

Do positive or negative discussions about specific tickers predict next-day returns?
Does the volume of online discussion correlate with abnormal trading volume?

3. How will you collect data and from where?

We will collect two main types of data, social media discussions and stock market data through public APIs or web scraping.

Social Media Data (Reddit): We will collect posts and comments between August 15 and November 15, 2025 (data spanning three months), via the Reddit PRAW API from r/stocks and r/wallstreetbets. The text, timestamps, and scores of posts mentioning stock tickers or cashtags will be extracted. We will then calculate the frequency of mentions per ticker and identify the top 5-8 most discussed stocks during this period.

Stock Market Data: We will use the free yfinance library (Yahoo Finance API) and download each selected stock's daily open, close, and volume data over the same period (August 15-November 15, 2025).

We will directly link social media activity to real market outcomes, so as to analyze whether daily sentiment and discussion intensity correlate with next-day stock price movements.

4. What analysis will you do and what visualizations will you create?

Data analysis:

We will perform descriptive statistics using Pandas to summarize key characteristics of sentiment scores and stock returns, including measures such as mean, standard deviation, and quantiles. This will be followed by correlation analysis to evaluate relationships between sentiment scores, discussion volume, stock returns, and trading volume.

Data visualization:

Firstly, we will draw a dual-axis time series chart of the sentiment score and the stock price trend to visually demonstrate the coordinated changes of the two in the time dimension. The correlation heat map will clearly present the intensity of the association between each variable in terms of color depth, highlighting the key relationship patterns. Through the yield box chart of emotion classification, the distribution differences of yields under different emotional states can be intuitively compared, including median, interquartile range and outlier characteristics. In addition, the scatter plot of discussion volume and trading volume, in combination with the trend line, will reveal the quantitative relationship pattern and density distribution between the two.