Project Topic Idea: Predicting Stock Prices via Twitter Sentiment Analysis

1. **Problem Statement**
   - Stock price predictions are volatile and highly sensitive to external factors. The rise of "meme stocks" such as Gamestop, which gained huge social media traction and astronomical gains and losses within mere hours, begs the question of the degree of influence social media has on stock prices.

   - To learn the degree of this influence, this project will focus on collecting and analyzing tweets about public stocks and use a predictive model to estimate the closing price of these stocks the following day.

2. **Significance of the Problem**
   - The project will help to not only predict stock prices, but to recognize the accuracy to which social media has influence on these security prices in terms of volatility and volume.

   - Here is an academic research paper I found after formulating this idea similar to what we hope to accomplish: Kordonis, John & Symeonidis, Symeon & Arampatzis, Avi. (2016). Stock Price Forecasting via Sentiment Analysis on Twitter. 10.1145/3003733.3003787. The methodology to analyze and process the sentiment of the Tweets will be a guideline to our project. Furthermore, the study did indeed find a correlation between the sentiment of tweets and stock prices, so we hope to express findings to a similar degree in our analysis.

3. **Potential Datasets**
   - The Twitter API allows us to search for recent tweets or stream tweets. This will help us collect textual data in the form of tweets to use for our sentiment analysis and feed our predictive model.

   - To collect historical and closing market data for our predictive model, we will use Yahoo Finance data, provided by existing APIs and libraries such as yfinance.

Dataset File

We prepared two datasets. One which contains historical financial data from yahoo finance using the yfinance API and another which contains tweets about the associated public companies.

Variable name in file	**Description (what the variable represents/means)[1]**	**Feature/ Outcome[2]**
Close	Close price of stock on a certain day	Feature
Volume	Amount of shares of stock traded on a certain day	Feature
Change %	Difference between open and close price of stock on a certain day	Feature
HL %	Difference between high and low price on a certain day	Feature
HPR	Total return including income from holding a share of stock from the previous day to the specified day	Feature
Market Capitalization	Total value of public equity in circulation for a company	Feature
Positive Sentiment Score %	Percent of n-grams (e.g. unigrams, bigrams - combinations of *two* consecutive sentiment-impactful words, trigrams) in a tweet deemed to be positive in nature (*todo from sentiment analysis model*)	Feature
Neutral Sentiment Score %	Percent of n-grams in a tweet deemed to be neutral in nature (*todo from sentiment analysis model*)	Feature
Negative Sentiment Score %	Percent of n-grams in a tweet deemed to be negative in nature (*todo from sentiment analysis model*)	Feature
Close	Close price of stock on a certain day (future)	Outcome
[1] Refer to dataset descriptions in sklearn [2] In the Feature/Outcome column, indicate whether the variable is a feature or outcome variable. You need to have at least one outcome variable and nine feature variables.		

Our dataset includes a set of feature variables listed above and one outcome variable which is the predicted close price of the stock in a future date that we can use for a supervised machine learning task. In order to do this, we will first construct a classifier to analyze the sentiment of our tweets and score what percent is deemed positive, neutral, and negative. This will use supervised learning and utilize methods from pre-existing sentiment lexicons customized by analyzing n-grams of our cleaned and filtered input. We will then use these results in our feature matrix for our second classifier which also uses supervised learning and historical market data to predict the future closing price of the stock based on tweets made in that interval.