Capstone Project 2 Proposal

Problem Statement: Analyze stock market movements using Twitter sentiment analysis to find the correlation between ”public sentiment” and ”market sentiment”

I. Background

The value of individual stocks often do not seem to reflect the fair value due to human error. Instead, the price stocks trade at seem to be determined more by the human perception of the stock. Behavioral economics states that the emotions and moods of individuals affect their decision making process. People and news outlets are constantly voicing their opinions about a variety of subjects, stocks included, on Twitter. Though a single tweet may not be significant, a large collection of them can provide data with valuable insight about the common opinion on a particular subject. Twitter, therefore, can be used to gauge the public sentiment and possibly predict stock price movements. The famous research paper by Bollen et al had performed Twitter sentiment analysis to predict price movement of the Dow Jones Industrial Average (DJIA). However, this study will focus on 4 individual stocks: Amazon ($AMZN), Google ($GOOGl), Disney ($DIS), and Netflix ($NFLX).

II. Impact

Applying sentiment analysis to stock movement forecasting has been prominent field due to the potential financial gains from it. Without a doubt, investment banking firms have done extensive research and built model based of the idea. Twitter API makes its data readily available to the public, so why shouldn't individual investors benefit from the wealth as well? By analyzing these trends and monitoring public opinion of companies, we can possibly build a predictive model to exploit market inefficiencies and anticipate changes in the market before they happen.

III. Data

A. **Twitter data**: The Twitter API will be used to pull tweets mentioning the stocks of interest. It is important to note that Twitter prevents users from pulling tweets past 7 days old with the standard API key. Sentiment Analysis "VADER" will be used to analyze the sentiment of the tweet. The data will include the following information:

1. Date and Time tweet was posted
2. Content of actual tweet
3. Follower count
4. Percentage of Negative sentiment
5. Percentage of Positive sentiment
6. Percentage of Neutral sentiment

B. **Historical Stock Data**: Quandl API will be used to retrieve historical data for the stocks of interest. It will only be pulled starting from the earliest tweet pulled. The data will include the following information:
    1. Date
    2. Open
    3. High
    4. Low
    5. Close
    6. Change
    7. Traded Volume
    8. Turnover
    9. Last Price of the Day
    10. Daily Traded Units
    11. Daily Turnover

IV. Outline

A. Retrieve data from Twitter and Quandl API
B. Perform sentiment analysis on the tweets using VANDER
C. Build a regression model that will predict the difference in stock price from the previous day