MAIS 202 Deliverable 1: Stock Market Predictor

Deepak Singh

I. Dataset

Being a student studying finance and computer science, I wanted to become the next billionaire on Wall Street, so I decided to implement machine learning with the stock market. I found this dataset on Kaggle comparing 2,000 days of financial news headlines with their respective industrial average stock prices. I am aiming to find an algorithm to classify the direction of the stock market according to financial news headlines.

i. Data Processing

Fortunately, the data is well processed in a single csv file with two variables: Down Jones Industrial Average (DIJA) price changes (increase vs decrease) and financial news headlines. The direction in price has already been labelled with dummy variables (0 and 1s) to respectably represent “decreasing” and “increasing” thus no further preprocessing is needed. The news article labels will need to be cleaned up in order to remove some unnecessary characters.

ii. Machine Learning Algorithm

The goal of this project is to predict the direction of stock price depending on the sentiment of the news article. To achieve this, I will implement a natural language processing algorithm in order to determine the positivity, negativity or neutrality of the headline which will lead me to determine the probability of the stock market increasing or decreasing. Subsequently, I will use a form of classification with three classes: positive, neutral, negative. For this project, I decided to use the Naïve Bayes model because of its low processing requirements compared to other classification models. This comes at the cost of disregarding the important sentiment implication of adjacent words. However, I feel that news headlines tend to use strong and straightforward vocabulary which makes these connections less significant compared to movie summaries or poems for example.

iii. Final Conceptualization

I will build a web application with React and JavaScript that will allow users to input current financial news article headlines and my model will output its resulting prediction for the Down Jones Industrial Average price versus the actual current price. To import the current price data, I plan on using web scrapping with Python or using a pre-made API. I will deploy the website on Heroku to host the web application.