**Data Mining Project**

**Stock market price prediction using news events**



## 

## 

## 

## Samyak Bagra 1610110333

## **Introduction**

For the project I plan to analyze the correlation between stock price and stock news data.

The stock market and news headlines datasets are likely to be correlated since national and international events have effects on economic outlook thus causing stock prices to fluctuate. I plan on finding if there are any correlations between the type of news event (war, presidential election outcomes) and changes to stock prices of companies in certain industries. To show these correlations I will apply exploratory data analysis (EDA) to the data, comparing variables such as names, places, and other key words in the headlines and seeing if there are any strong connections to stock prices.

## **Hypothesis and Questions**

Will the positive news sentiment increase the stock price, and will a negative news sentiment decrease the stock price?

Does news at least have a slight effect on stock prices?

I believe it does, but it doesn't matter if it's positive or negative. Like with the recent Coronavirus headlines, even though the stock is dropping a lot, there are still up days even when the news is negative.

## **Datasets used**

* Stock News headlines dataset (Downloaded): [https://www.kaggle.com/aaron7sun/stocknew](https://www.kaggle.com/aaron7sun/stocknews)s

There are two channels of data provided in this dataset:

1. News data: Crawled historical news headlines from [Reddit WorldNews Channel](https://www.reddit.com/r/worldnews?hl) (/r/worldnews). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.  
    (Range: 2008-06-08 to 2016-07-01)
2. Stock data: Dow Jones Industrial Average (DJIA).  
    (Range: 2008-08-08 to 2016-07-01)

**Combined*News*DJIA.csv**:  
 This combined dataset has 27 columns. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25".

This is a binary classification. Hence, there are only two labels:

“1" when DJIA Adj Close value rose or stayed as the same;

"0" when DJIA Adj Close value decreased.

## **EDA Process**

I first cleaned the data by deleting any null rows and converting all the headlines to strings and all the numbers to floats. The first analysis I did was seeing the frequencies of all the values within each column. I then compared how different aspects of the news headlines sentiment had to do with an up or a down day for the stock. I then did two pair plots. The first with the raw data and the second with the regression data. The regression data was more readable so it was easier to analyze. Finally I did a parallel coordinates plot comparing the positivity, negativity, and neutrality of the news headline with whether it was an up or a down day for the stock. The analysis I did didn't reveal a lot, but I was able to get a small amount of useful information.

## **Summary of Findings**

Something that was revealed within the data was that there isn't that much of a correlation between the positivity/negativity of the news and whether it is an up day or a down day. There is a slight edge in the market, however, when the news is negative. This slight edge being that the stock is slightly more likely to have an up day when the news is negative. I conclude from this that the stock has been down for a couple of days due to a couple days of bad news thus leading to a "discounted stock" that is more appealing to buy to investors. This increase in the amount of stocks bought on that day would thus lead to an increase in the price of the stock. As for positive news, investors usually sell their stocks at a high, or when they think the stock is above its value. This then leads to a drop in the price of the stock due to a high frequency of sells. All in all, I believe that the news slightly affects the stock prices, but not that much. Many of the stock ups and downs are due to a majority of people just wanting to buy or sell based on the price. The stock market is just based on the amount of stocks bought and sold, and not of news so it is logical to deduce that the news has no effect unless it directly has to do with that specific stock.