**Problem Statement**

To build a classification model using Natural Language Processing techniques to identify, based on bunch of news headlines, if a stock price has increased (class 1) or decreased (class 0).

**Architecture**



**Data Details**

1. The data in consideration is a combination of world news and stock price shifts available in Kaggle.

<https://www.kaggle.com/aaron7sun/stocknews>

1. There are 25 columns of top news headlines for each day in the dataframe.
2. The data ranges from 2008 to 2016.
3. Labels are based on Dow Jones industrial Average Stock Index.

Class 0 🡪 The stock price increased.

Class 1 🡪 The stock price decreased.

**Steps Implemented in Model**

**Step 1:** Downloaded the data from Kaggle from the link given above.

**Step 2:** Imported the data into a pandas dataframe and performed data cleaning on all features. The steps include renaming columns, bringing all text to lower case, deleting redundant news, changing datatypes, and replacing nulls with blank space.

**Step 3:** Concatenated all text columns (all 25 news headlines) and applied stemming to it and removed stop words and special characters. NLTK library’s PorterStemmer() was used and output is stored in a list.

**Step 4:** Applied Bag of Words and TFIDF to the data in separate ipynb files.

**Step5:** Performed hyperparameter tuning on the data using the algorithm given below

* Random\_Forest
* naive\_bayes\_multinomial
* Decision\_Tree
* SVM

**Files to run the model**

1. Stock Sentiment Analysis - Bag of Words.ipynb
2. Stock Sentiment Analysis - TFIDF.ipynb
3. Stock.csv

**Observation**

Decision Tree Classifier seems to win with highest accuracy; however the accuracy is not very high because the Bag of Words and TFIDF algorithm does not work well with semantics of the data. For this sentiment analysis, the news semantics are very important to change the emotion and opinion of the traders to impact the stock market. RNN and Word 2Vec is a better algorithm here and will be further explored.