**Internship Project Document (to be modified as we progress)**

**Initial Idea: Predicting the Stock Market(up/down) with News Articles**

**Description**:

Since stock market depends on the market and investor’s sentiments, we can use the latest news to predict whether stock price for any certain company is going up or down. A very basic nlp based approach can be seen here : <https://nlp.stanford.edu/courses/cs224n/2007/fp/timmonsr-kylee84.pdf> .

**Front End**:

The project can be deployed as either a mobile app or webapp. We can also add a chatbot for user interaction.

The user can give two inputs:

* Name of the company (from an available list).
* Number of past days(1-3days) to consider for news articles.
* News Article Sources (from an available list).

The output of ‘predict’ action will be:

* The up/down predicted probability for stock price.
* Past trend as graph.
* Top-x features influencing the final prediction.

The Chatbot can handle below interactions:

* Take user inputs and interactively return the output as mentioned above.

**Data Required**:

For training:

* At initial stage, we will only focus on selected companies (50 companies listed in Nifty50).
  + <https://www.moneycontrol.com/stocks/marketstats/nse-gainer/nifty-50_9/>
* Daily stock price(open/close/min/max/volume) for past one year (or 6 months).
* Financial News Article for past one year related to these 50 companies.
* People’s opinions/comments for past 1-2 day.

For prediction:

* Past few day’s news scraped from the same news sources used in training.
* Past few day’s stock prices (depends on our final features in model).

Data Sources:

* Yahoo Finance for daily prices of each company.
* NSE Bhavcopy.
* Moneycontrol or any other website for news/people’s opinions.

**Approach**:

For model training:

* Target Variable(binary): up/down (based on ‘open’ and ‘close’ prices).
* Input Features: past high/low/volumes traded and text from news.
* Model Training: We can start with some basic word tokenization/tf-idf/word2vec/lstm techniques and then proceed to transfer learning models (like BERT).
* We can also see if multiple stock prices are correlated in some way.
  + <https://github.com/imhgchoi/ARIMA-LSTM-hybrid-corrcoef-predict>.
* We can also explore the possibility of combining time series(ARIMA/lstm/varmax/prophet) and text classification models for a better result.
  + <https://www.kaggle.com/janiobachmann/s-p-500-time-series-forecasting-with-prophet>