Stock Market Prediction Using Sentiment Analysis

Group Members:

Gaurav Mahadik - 801080505.
Pratik Parekh - 801076521.
Sakshat Surve - 801080533.
Sumedh Joglekar - 801084301.

Introduction:

Stock Market prediction has been an active interest in the world for quite a few years now. Stocks are majorly based on human behavior or reactions to certain changes. Nowadays most of the activity of humans can be tracked on social media. And any events that occur affect the human behavior and in turn the stock market. Thus, by extracting patterns and analyzing this data we try to predict the stocks of certain company.

We are doing this using certain Machine Learning algorithms and Neural Networks. Data Extraction is done from various sources and methods such as web crawling from various websites, web scraping for news and stock data from World Trading Data and collecting tweets from Twitter.

Problem Statement:

1. The use of social media and other sources to assess the market sentiment and predict the changes in the stocks of a company.
2. Classification of the polarity of a given phrase or sentence to determine whether the opinion is positive or negative.
3. Building a machine learning model to predict the sentiments across different platforms and finding the degree of correlation between the sentiment and the behavior of the stock price of a company.

Dataset:

- We will be using more than one sources as the dataset for predicting the behavior of stocks according to the sentiments identified in the dataset.
- The multiple data sources will ensure the randomness required, that is required in order to learn the different sentiments which will be used to predict the behavior of the stocks.

We will collect data by:
1. Web crawling for ticker list from multiple websites such as NASDAQ, NYSE.
2. Web scraping for news from multiple websites such as Reuters and Bloomberg.
3. Extraction of stock price data from WorldTradingData, Alpha Vantage and Yahoo API.
4. Collecting data from tweets in order to determine the sentiment associated with it.

Approach:

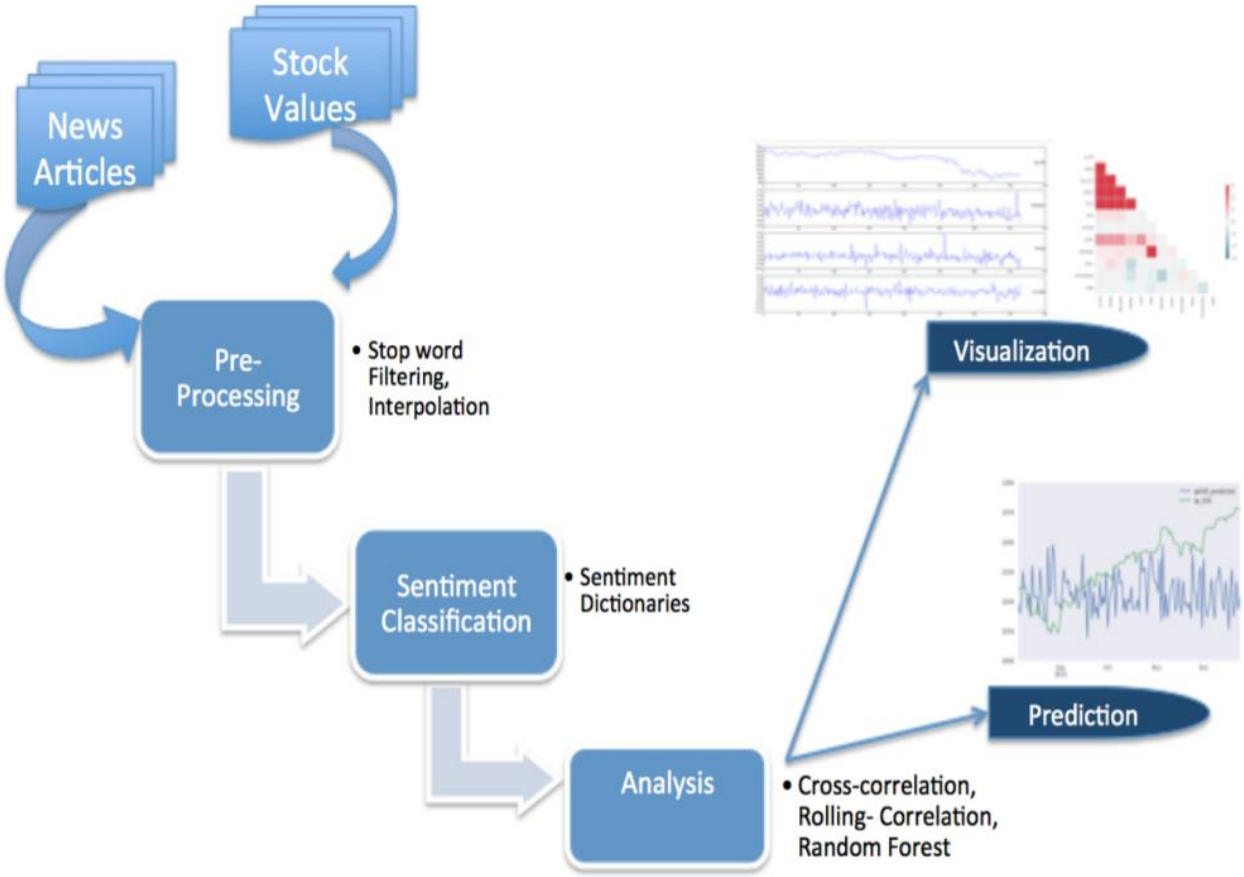
Below are the main steps which will be involved in our project:-

1. Data Collection:
    - We will be needing the stock price data and also the data related to the sentiments.
    - To get the latest stock price data, we will be using web crawling on different websites.
    - First we will crawl the ticker list to get the details of the companies.
    - Once we get the company list then we will crawl the news with the help of BeautifulSoup to get the sentiment data.
    - To extract real-time stock price data, we are planning to use available APIs such as WorldTradingData /Alpha Vantage / Yahoo API.

2. Data Cleaning and Preprocessing:
    - The collected data needs to be cleaned i.e. some unwanted elements of the data need to be removed and formatted in a desirable format which can be fed to the model for training.
    - Tokenization will be performed to break down the sentences into individual words in order to generate the aggregate score of the sentence so that the sentence can be classified as positive or negative.
    - Stop words, punctuation, numerical, special characters, hashtags in case of tweets will be removed from the new headline and tweet data.
    - Lemmatization will be performed to cut down the words to their root words so that there is no repetition of words. Splitting can be done by using spaces. Then the words are converted into the lower cases.
    - Once the tokenization/lemmatization is performed on the data, data needs to be converted into the format which will be suitable for further analysis. And hence we will convert the data into vector format.

3. Training :-
    - Once the data is ready to be used in the model, we will be using the pre-processed data for the training. The training step will apply the convolution neural network on the training data.
    - Convolution neural network will predict the stock price movement based on the selected parameters

4. Prediction:-
    - Whenever the new news/tweets appears related to the specific company, then the based on the train model, the score of an event will be calculated.
    - Based on the calculated score, prediction will be made regarding the trading strategy.


News Articles
Stock Values
Pre-Processing
• Stop word Filtering, Interpolation
Sentiment Classification
• Sentiment Dictionaries
Analysis
• Cross-correlation, Rolling- Correlation, Random Forest
Visualization
Prediction

Literature Survey:

- **Using Structured Events to Predict Stock Price Movement: An Empirical Investigation:** As news events affect human decisions and the volatility of stock prices is influenced by human trading, it is reasonable to say that events can influence the stock market. In this paper, they have presented a framework for event-based stock price movement prediction and extracted structured events from large-scale news based on Open Information Extraction technology and employed both linear and nonlinear models to empirically investigate the complex relationships between events and the stock market. They have used publicly available financial news from Reuters and Bloomberg over the period from October 2006 to November 2013. In this time span witnesses a severe economic downturn in 2007- 2010, followed by a modest recovery in 2011- 2013.

  http://www.aclweb.org/anthology/D14-1148.

- **Stock Prediction Using Twitter Sentiment Analysis:** In this paper, there is a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decision-making process, thus, leading to a direct correlation between public sentiment and market sentiment. This method is only restricted to people that use Twitter or more specifically English speaking users on Twitter, thus not giving us the real sentiment analysis. Thus, we are also using other methods such as Information extraction from News, in our project.

  http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

- **Stock Prediction Using Event-Based Sentiment Analysis**:- This paper describes a method to predict future stock market movement using twitter sentiment. The major part of this implementation is automatically generating the training data based on the events related to stock market. This training data is used to build a high precision model to classify the tweet sentiment. Using this information, future stock market movements are predicted which eventually can be used for trading strategy. In this approach, tweets of all stock constituents were collected using Twitter search API.

  https://ieeexplore.ieee.org/document/6690034

- **Deep Learning for Event-Driven Stock Prediction:-** This paper describes the deep learning method for event driven stock market prediction. Events are extracted with the use of Open IE technology and the dependency parsing. They have also used the neural tensor network to convert the word embedding into the event embedding. With the use of three categorized events - long term, mid term and short term change in the stock price was predicted based on the framework of CNN. The model developed in this paper considered that the influence of monthly news and weekly news are less effective as compared to the latest news on the stock prediction. Experimental results showed that event embeddings-based document representations are better than discrete events-based methods. As

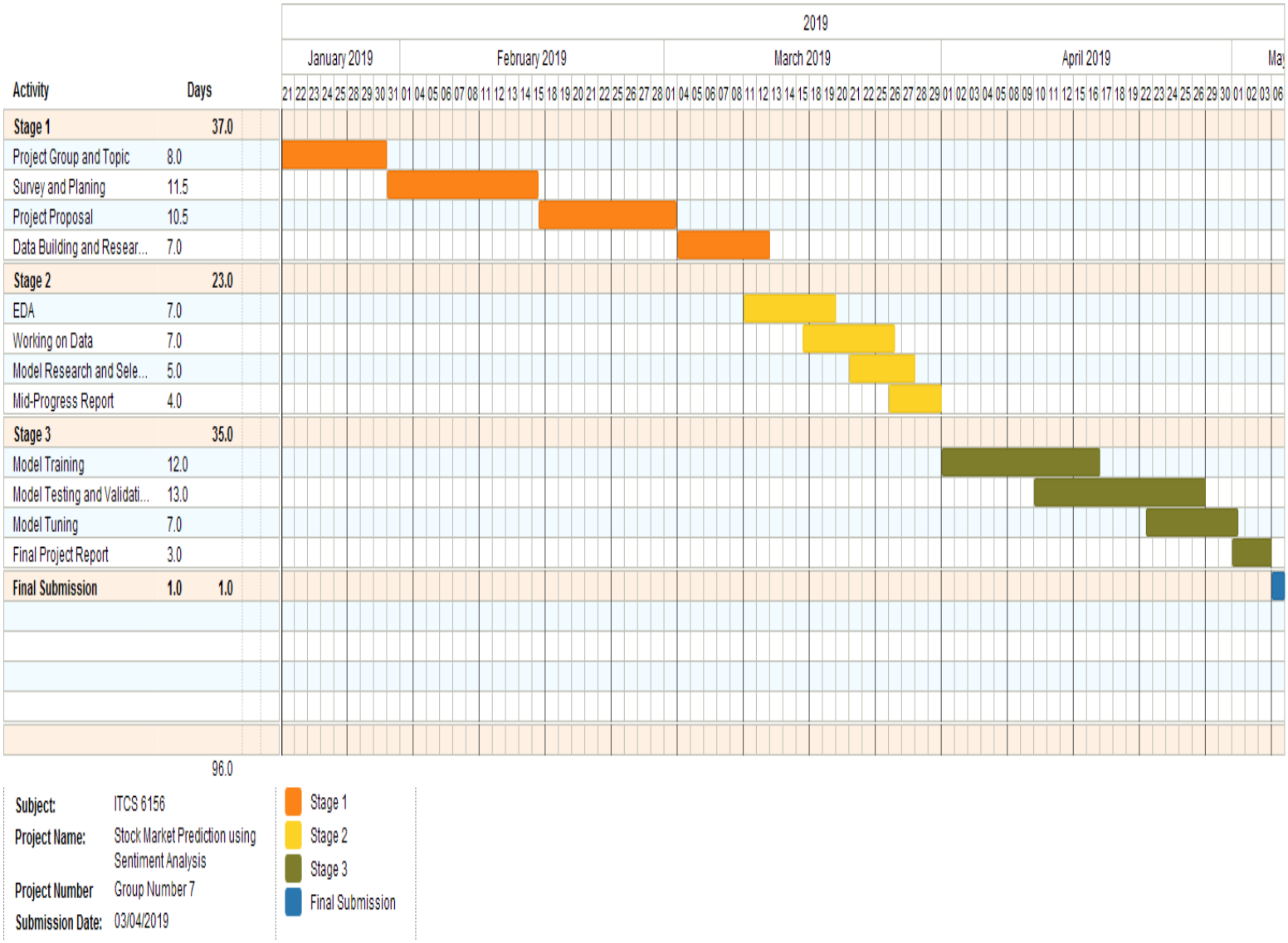
the model developed in this paper shows better results in the prediction of the stock prices and hence the same can be used in our project while doing the prediction.

https://www.ijcai.org/Proceedings/15/Papers/329.pdf

- **Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation:-**This paper describes the different approach to predict the stock market prices. This approach makes use of Twitter data to gather the sentiments and then based on the positive or negative sentiment of tweets (searched with the use of specific keyword), the stock price was predicted. We found that results obtained using the methods mentioned in this paper can be improved by large time granularity and more data. Hence we will be implementing sentiment analysis on a large scale so that stock prediction will be more accurate. Along with the news data, twitter data can also be used in our model while calculating the sentiment score and this approach may be useful to achieve good accuracy.

  https://apps.cs.utexas.edu/tech_reports/reports/tr/TR-2124.pdf

Timeline:


2019
January 2019
February 2019
March 2019
April 2019
Activity
Days
Stage 1 37.0
Project Group and Topic 8.0
Survey and Planing 11.5
Project Proposal 10.5
Data Building and Resear... 7.0
Stage 2 23.0
EDA 7.0
Working on Data 7.0
Model Research and Sele... 5.0
Mid-Progress Report 4.0
Stage 3 35.0
Model Training 12.0
Model Testing and Validati... 13.0
Model Tuning 7.0
Final Project Report 3.0
Final Submission 1.0 1.0
96.0
Subject: ITCS 6156
Project Name: Stock Market Prediction using Sentiment Analysis
Project Number Group Number 7
Submission Date: 03/04/2019
Stage 1
Stage 2
Stage 3
Final Submission

Roles of Each Team Member:

Sr No	Tasks Performed	Gaurav	Pratik	Sakshat	Sumedh
1.	Project Group and Topic	✓	✓	✓	✓
2.	Survey and Planning	✓	✓	✓	✓
3.	Project Proposal	✓	✓	✓	✓
4.	Data Building and Research		✓	✓	✓
5.	Exploratory Data Analysis	✓	✓		✓
6.	Working on Data	✓	✓		✓
7.	Model Research and Selection	✓	✓	✓	
8.	Mid-Progress Report	✓	✓	✓	✓
9.	Model Training	✓		✓	
10.	Model Testing and Validation	✓		✓	✓
11.	Model Tuning		✓	✓	✓
12.	Final Project Report	✓	✓	✓	✓

Questions answered by our Project:

- ➢ Can Machine learning algorithms be used for predicting stocks in real?
- ➢ Can Stocks be predicted by people's behavior or events happening in the world?
- ➢ Whether a Tweet about certain Stock is negative or positive?
- ➢ After analyzing results of the model, what will be the trading strategy based on the sentiments?

Things we will learn:

We are going to use more than one machine learning algorithms. In these, we learn about how a machine learning algorithm can be used to predict stocks on certain parameters. We will also learn how to preprocess and clean data. The neural network that we will use will help us understand, how the events affect the stocks. We are also going to perform certain web scraping techniques such as Beautiful Soup, that will help us in Information Extraction. Stemming, Lemmatization and Tokenization are among the data cleaning terms that we are going to learn. The model used on training data will later be used on the testing data to predict the stocks.

Is the Idea Novel:

Not entirely. Stock prediction is a very emerging domain in machine learning and many new models/methods are getting introduced every day to improve the efficiency of the prediction. But for our project, we are planning to use a neural network to build a model on sentiment events and make more accurate predictions on stock price movements.

References:

1. Using Structured Events to Predict Stock Price Movement, An Empirical Investigation, Referenced from - http://www.aclweb.org/anthology/D14-1148
2. Stock Prediction Using Twitter Sentiment Analysis , Referenced from - http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf
3. Stock Prediction Using Event-Based Sentiment Analysis, Referenced from - https://ieeexplore.ieee.org/document/6690034
4. Deep Learning for Event-Driven Stock Prediction, Referenced from - https://www.ijcai.org/Proceedings/15/Papers/329.pdf
5. Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, Referenced from - https://apps.cs.utexas.edu/tech_reports/reports/tr/TR-2124.pdf
6. Deep Learning for event driven stock prediction , Referenced from - https://medium.com/@wenchen.li/deep-learning-for-event-driven-stock-prediction-ab783b322f19
7. Real Time Market Data Referenced from - https://www.worldtradingdata.com/documentation#stock-and-index-real-time
8. Real Time News data, Referenced from - https://www.reuters.com/