# BLUEPRINT

## Predicting market volatility and building shot-term trading strategies using data from Reddit's WallStreetBets

-Title

**Submitted By:**
**Bidhya Pokharel**

**22nd August, 2021**

## 1. Requirement (Business) :

Stock prices are very dynamic and susceptible to quick changes because of mix of various known and unknown parameters. But at this time, looking at the recent histories of stock market we can conclude that News or Social Medias or Investors/Leaders tweets has become one of the major part to fluctuate the stock prices by billion within a second.

So, considering this fact our aim is to predict the future price of security through analyzing posts on WSB (WallStreetBets: subreddit where participants discuss stock and option trading). The long term goal of the project is to build a trading bot that intelligently executes short-term trades based on insights gleaned from various different news portals or sources.

## 2. Data Acquisition :

As per project requirement, the data from the news portals or social media sites which has lot of stock enthusiast, Customers, Investors, Companies and so on is needed. So, the data from Reddit's WallStreetBets can be used. For this project, two different datasets will be used.

a. SPX file which contains lot of 'High', 'Low', 'Close', 'Open', 'Volume', so we'll trim it out as per the Kaggle file named SP500. Hence from this file we'll get four columns:

- i. Open
- ii. Close
- iii. High
- iv. Low
- v. Volume
- vi. Date

b. JSON file that contains the body of WSB post. Since, this file is very large we'll take out only

- i. Body
- ii. Date
- iii. Score

We'll also create the **target variable** using open and close value. Where our condition will be, if yesterday's closing price is smaller than today's closing price then our value will be 1 similiarly vice versa.

## 3. Data Processing :

Data Processing is one the most important part of Data Science projects. Low Quality data can affect our model negatively so before going for further steps we should clean our data and transform/parse the data input into the format that our ML algorithm can understand. Python libraries Pandas and Numpy can be used for this purpose.

So, first we'll perform **Sentiment Analysis** in the body column that we had received from JSON file. From where we'll get the 'Positive', 'Negative' and 'Neutral' value. Then we'll merge this values with previously trimmed CSV files on the basis of Date. Hence, after which we'll perform data cleaning involving following steps:

i. Impute/Remove missing values or Null values (NaN)
ii. Remove unnecessary and corrupted data.
iii. Date/Text parsing if required.

## 4. Data Exploration:

After getting the data cleaned, exploration of data is necessary to understand the patterns in data or to retrieve useful insights/relation.

a. Exploratory Data Analytics (EDA):

We can find the correlation among the dependent and independent features so that we can explore which feature is more important (Feature Selection Method) via different methods like:

i. Matplotlib plots like Histogram, Heatmap, etc.
ii. Metrics like MAE, MSE and so on.

b. Feature Engineering:

We can encode our categorical data if necessary using different classifiers/labels like Label Encoder, Binary encoder and so on. We will find the important features for our model.

## 5. Model Planning

With the optimal data features that we received from previous steps or as mentioned in project paper provided:

1. post_score
2. total_comms_num
3. len_of_post
4. positive_sentiment
5. neutral_sentiment
6. negative_sentiment
7. bias (column of 1’s)

Since, the project aim is to predict market volatility, the project as be categorized as a Classification problem for which various following algorithms can be used:

1. Logistic Regression
2. Random Forest
3. Decision Tree
4. Naive Bayes Classifier Algorithm
5. Support Vector Machine Algorithm

## 6. Model Building:

Its not necessary to be limited to ML algorithm only, Deep learning can also be used for this problem because Neural Networks not only have the ability to discover patters in non-linear and chaotic systems but also offer the ability to predict market directions more accurately than most other current techniques. So, basically following experiments can be done:

a. Logistic Regression:

Explore the above mentioned algorithm by dividing the provided data sets into training and testing set, and build the model using the best algorithms or the algorithm which gives higher accuracy. Hyper tuning can be done to reach the best model of our project.  So, basically as per the project paper provided Logistic regression can be used to for higher accuracy.

b. Neural Network:

Create, compile and fit a model using specific metrics, epochs, activation and optimization function. If the model output is not as expected then we can further evaluate it by increasing the number of layers and epochs, changing the activation and optimization function. Repeat the evaluation until we get the improved or best model. Hence finally we can save our improved trained model.

## 6. Model Deployment

After building the model, we can deploy it using different cloud deployment model like Azure, AWS, etc so that we can reach to all the possible users. So, I’m looking forward to deploy the model using AWS EC2.

## 7. Conclusion :

From the experiments, we can take the model which gives us higher accuracy. Figure out, if there is correlation among the dependent and independent features. We can conclude if the WallStreetsBets really/how deeply impact the stock fluctuation.

### 8. Future Work:

Data used for this project is from WallStreetsBets only, so if we’ll be able to expand our data from other related sources as well, it would be really great.