Stock Market Prediction using Machine Learning and Natural Language Processing

Aditya Prasanna[1], Dheemant Surendra[2], Eshika Anil[3], Hemalatha R[4], Mrs. Shubha T V[5]

[1]*Student, Dept. of Computer Science Engineering, SJB Institute of Technology, Karnataka, India*
[2]*Student, Dept. of Computer Science Engineering, SJB Institute of Technology, Karnataka, India*
[3]*Student, Dept. of Computer Science Engineering, SJB Institute of Technology, Karnataka, India*
[4]*Student, Dept. of Computer Science Engineering, SJB Institute of Technology, Karnataka, India*
[5]*Assistant Professor, Dept. of Computer Science Engineering, SJB Institute of Technology, Karnataka, India*

---***---

Abstract - *Stock value anticipating is a mainstream and significant subject in monetary and scholastic examinations. Monetary markets are profoundly unpredictable and produce colossal measures of information every day. It is the most well known budgetary market instrument and its worth changes rapidly. In any case, the financial exchange is impacted by numerous elements, for example, political occasions, monetary conditions and brokers' desire. Numerous techniques like specialized examination, major investigation, time arrangement investigation and factual examination, and so forth. In this project we attempt to implement an Artificial Neural Network and Natural Language Processing using Sentimental Analysis to predict stock market prices, returns, and stock modelling, and the most frequently used methodology is the Backpropagation algorithm. The framework has a graphical UI and capacities as an independent application. The framework can be stretched out to dissect multivariate time arrangement information and import crude dataset straightforwardly. The advancement of an online application has been considered to improve the ease of use and ease of use of the master framework.*

***Key Words*: Stock price, Machine Learning, Artificial Neural Network, Natural Language Processing, Sentimental Analysis.**

1. INTRODUCTION

A stock (otherwise called shares all the more usually) when all is said in done speaks to proprietorship asserts on business by a specific individual or a gathering of individuals. Individuals will in general purchase those stocks whose costs are relied upon to ascend sooner rather than later. The vulnerability in the securities exchange abstain individuals from putting resources into stocks. Therefore, there is a need to precisely foresee the financial exchange which can be utilized in a genuine scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning modelling and predicting the variable stock market. The two financial specialists and industry are engaged with securities exchange and needs to know whether some stock will rise or fall over certain timeframe. It depends on the idea of interest and supply. On the off chance that the interest for an organization's stock is higher, at that point the organization share cost increments and in the event that the interest for organization's stock is low, at that point the organization share value decline. Examination of stocks utilizing information digging will be helpful for new financial specialists to put resources into securities exchange dependent on the different variables considered by the product. Investment is a current commitment in the expectation of getting greater resources in the future. In addition, the purpose of a person doing an investment is to improve their welfare. A media for investment activities is named as capital market. The term capital market is often understood as a stock market and the stock is often referred to as common stock information on stock market performance is often summarized in a stock market index. The stock market index that often called stock price index is an indicator that reflects the performance of stocks in the market.

1.1 EXISTING SYSTEMS

The existing system majorly utilizes use of Machine Learning in Stock Prediction. The concept of Support Vector Machines (SVM) has advanced features that are reflected in their good generalization capacity and fast computation. SVMs can be utilized to perform Linear Regression on past stock information to foresee the end costs utilizing Time arrangement anticipating and other enhancement calculations. The most important of these is the Efficient Market Hypothesis (EMH), the hypothesis says that in an efficient market, stock market prices fully reflect available information about the market furthermore, its constituents and along these lines any chance of acquiring overabundance benefit stops to exist. The utilization of man-made brainpower procedures to anticipate the costs of the stock is an expanding pattern. The yield differs for every system regardless of whether similar informational collection is being applied. Due to the vast number of options available, there can be n number of ways on how to predict the price of the stock, but all methods don't work the same way. In the referred to paper the stock value forecast has been completed by utilizing the arbitrary woodland calculation is being utilized to anticipate the cost of the stock utilizing budgetary proportions structure the past quarter. This is only one perspective on issue by moving toward it utilizing a prescient model, utilizing the arbitrary timberland to foresee the future cost of the stock from authentic information.

1.2 MOTIVATION

The motivated idea is that, if we know all information about today's stock trading (of all specific traders), the price is predictable. Thus, if we can obtain just a partial information, we can expect to improve the current prediction lot. In this way, our inspiration is to structure an open help fusing authentic information and clients expectations to make a more grounded model that will profit everyone. Thus, our motivation is to design a public service incorporating historical data and users predictions to make a stronger model that will benefit everyone. These anticipated and examined information can be seen by individual to know the budgetary status of organizations and their correlations. Organization and industry can utilize it to breakdown their constraint and upgrade their stock worth. It tends to be valuable to even scientists, stock agents, showcase creators, government and general individuals.

1.3 PROPOSED SYSTEM

The proposed model is a promising prescient system for exceptionally non-straight time arrangement, whose examples are hard to catch by conventional models. The framework can be stretched out to dissect multivariate time arrangement information and import crude dataset straightforwardly. In our proposed framework, in light of the contribution of the client we proposed him/her the arrangement of stocks that is appropriate for their decision.

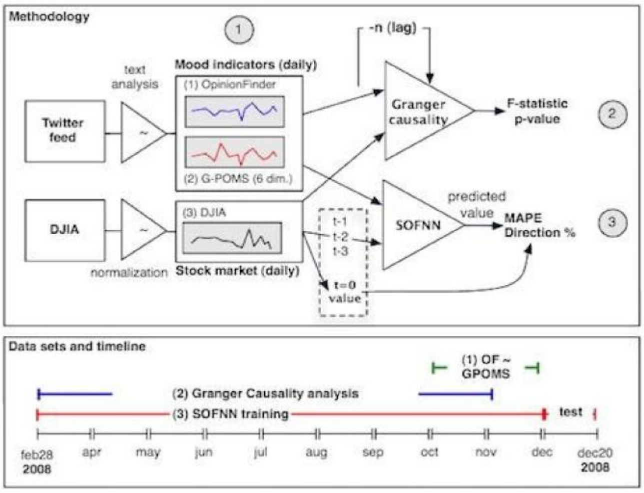

Fig -1: Stock Market Predictor Model

To set up this set, we considered distinctive central variables like EPS, P/E, beta, and co-connection, standard deviation and so on that we have caught, removed or inferred straightforwardly or in a roundabout way during data parsing process. Consequently, the choice technique of stock really depends how much hazard a financial specialist is happy to take in the market. However, these essential factors are just assistance for us to screen out and set up a protected rundown of stocks relying upon hazard factor setting yet to get the suitable time of venture on these stocks, we have to consider value pattern related variables which alludes to specialized examination.

2. DESIGN AND ARCHITECTURE

This project focuses on the usage of ingenious machine learning algorithms to achieve an implementation of the algorithm that gives maximum accuracy. The stock market depends on a lot of factors like the emotions of the CEOs and the emotion of the investors in general. Natural Language Processing is used to focus on capturing hints from the news articles, tweets, and personal statements made by the people linked with the stock market.

A. Feature Selection

A feature selection algorithm can be viewed as the blend of a quest method for proposing new element subsets, alongside an assessment measure which scores the distinctive element subsets. The least difficult calculation is to test every conceivable subset of highlights finding the one which limits the blunder rate.

B. Support Vector Machines

Support vector machines are supervised learning models with related learning calculations that investigate information utilized for order and relapse examination. Given a lot of preparing models, each set apart as having a place with either of two classifications, a SVM preparing calculation fabricates a model that doles out new guides to one classification or the other, making it a non-probabilistic paired direct classifier.

C. Decision Trees

A decision tree is a choice help apparatus that utilizes a tree-like model of choices and their potential results, including chance occasion results, asset expenses, and utility. Decision trees are usually utilized in tasks examine, explicitly in choice investigation, to help distinguish a procedure well on the way to arrive at an objective, but on the other hand are a famous device in AI.

D. Naïve Bayes

Naive Bayes classifiers are a group of basic "probabilistic classifiers" in view of applying Bayes' hypothesis with solid freedom suspicions between the highlights. They are among the most straightforward Bayesian system models. Naive Bayes classifiers are profoundly adaptable, requiring various parameters straight in the quantity of factors (highlights/indicators) in a learning issue.

E. Natural Language Processing

Natural Language Processing is a subfield of etymology, software engineering, data building, and man-made brainpower worried about the communications among PCs and human (common) dialects, specifically how to program PCs to process and break down a lot of regular language information. Natural Language Processing is extensively characterized as the programmed control of characteristic language, similar to discourse and content, by programming.

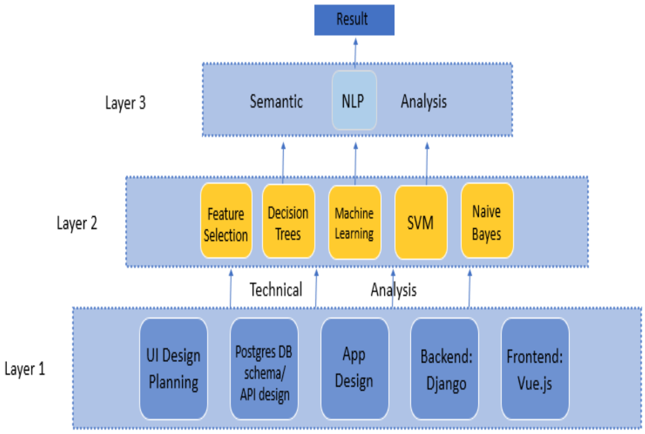

Fig -2: Various Layers of the Model

3. METHODOLOGY

We arrive at the most crucial aspect of the project pipeline. The different components of the project will be individual services interacting with each other through a variety of REST APIs and WebSockets. This is widely known as the revered Services Oriented Architecture or SOA. Here, the different services interacting with each other are the frontend (client), the backend (server), the WebSocket. The database used is PostGreSQL. The client interacts with the various services through the backend. The backend will be running on Django, the python based framework and the frontend will use Vue.js, the javascript based framework. The REST APIs we will be created using Django Rest Framework and the WebSockets will be built using Django channels. REST is an architectural paradigm which primarily uses HTTP 1.1 for unidirectional requests and responses. In contrast, a WebSocket is a protocol of its own, which is used for bidirectional streaming of bytes between the client and server. It is extremely fast, efficient and does not cause network congestion. We will be using Websockets to stream live stock prices (Last Traded Price). The Last traded price is the last price at which the asset was bought at the exchange. This price will be the pivotal element of the entire prediction model as the predicted outputs will be compared to this LTP for signal generation. In order to generate these signals, our machine learning models must be trained on historical prices first for predictions. These historical prices are widely available and can be collected from the internet. For the purpose of this project, we will be collecting a large data set worth 10 years of stock prices. The Last traded price will be collected from a commercial API from a broker called Zerodha through their kite.trade websocket. Lastly, our sentimental analysis model will require a fairly large dataset for accurate predictions and thus, we will be acquiring this by scraping tweets (from Twitter). Tweets are found to be the most common way of expressing sentiment in the financial services industry.

4. IMPLEMENTATION

Information Flow for Sentimental examination on twitter:

Stage 1: Tweepy: tweepy is the python customer for the official Twitter API. Introduce it utilizing following pip order.

Stage 2: TextBlob: textblob is the python library for handling printed information. Introduce it utilizing following pip order: Also, we have to introduce some NLTK corpora.

Stage 3: Verification: In request to get tweets through Twitter API, one needs to enroll an App through their twitter account. Follow these means for the equivalent:

1. Open app.twitter.com and snap the catch: 'Make New App'
2. Fill the application subtleties. You can leave the callback url field unfilled. Once the application is made, you will be diverted to the application page.
3. Open the 'Keys and Access Tokens' tab.
4. Copy 'Customer Key', 'Shopper Secret', 'Access token' and 'Access Token Secret'.

Stage 4: Execution: We follow these 3 significant strides in our program:

1. Authorize twitter API customer.
2. Make a GET solicitation to Twitter API to get tweets for a specific question.
3. Parse the tweets. Order each tweet as positive, negative or nonpartisan.
4. First of all, we make a TwitterClient class. This class contains all the strategies to communicate with Twitter API and parsing tweets. We use init capacity to deal with the validation of API customer. In get tweets work, we use: to call the Twitter API to get tweets. In get tweet estimation we use textblob module.
5. TextBlob is really a significant level library worked over top of NLTK library. First we call clean tweet strategy to evacuate joins, exceptional characters, and so forth from the tweet utilizing some basic regex. Then, as we pass tweet to make a TextBlob object, following handling is done over content by textblob library:
6. Tokenize the tweet ,i.e split words from assemblage of content.
7. Remove stopwords from the tokens.(stopwords are the regularly utilized words which are superfluous in content investigation as am I, you, are, and so forth.)
8. Do POS(grammatical feature) labeling of the tokens and select just huge highlights/tokens like descriptors, verb modifiers, and so on

9. Pass the tokens to a conclusion classifier which arranges the tweet estimation as positive, negative or nonpartisan by appointing it an extremity between - 1.0 to 1.0.
10. TextBlob utilizes a Movies Reviews dataset in which surveys have just been named as positive or negative.
11. Positive and negative highlights are separated from every positive and negative audit individually.
12. Training information currently comprises of marked positive and negative highlights. This information is prepared on a Naive Bayes Classifier.
13. Then, we use sentiment.polarity technique for TextBlob class to get the extremity of tweet between - 1 to 1.

5. RESULTS

As shown in the implementation section, we have taken the combined weighted average of the accuracy of the machine learning algorithms in stock trades and have applied sentimental analysis of natural language processing, this resulted in improved accuracy which yielded higher results compared to the last proven results.

6. CONCLUSIONS

Through this undertaking, it helped us comprehend the nuts and bolts of Natural Language Processing. Despite the fact that you can't wager your cash on the stock from this task, this work can be treated as strong comprehension of the essentials of Natural Language Processing. Utilizing a similar model for various content information is likewise attainable. It was intriguing to think about how to go from content information to vectors of numbers and applying Machine learning strategies that can assist with affecting the securities exchange of an organization. It helped us an increase more extensive feeling of the intensity of NLP in different applications. We likewise went from utilizing an accessible informational index to scratching our own information which made this venture somewhat more fascinating and testing. To get more experiences on which model to utilize and how to develop them, we learnt by perusing research papers and utilization of scikit figure out how to manufacture our models.

REFERENCES

[1] R. Wilson and R. Sharda, "Bankruptcy prediction using neural networks", Decision Support Systems, Vol. 11, No. 5, pp. 545-557, 1994.

[2] Atsalakis, G. S., Dimitrakakis, E. M., & Zopounidis, C. D. (2011). Elliott wave theory and neuro-fuzzy systems, in stock market prediction: The WASP system. Expert Systems with Applications, 38, 9196–9206.

[3] K. Tsai and J. Wang, "External technology sourcing and innovation performance in LMT sectors", Research Policy, Vol. 38, No. 3, pp. 518-526, 2009.

[4] K. Han and J. Kim, "Genetic quantum algorithm and its application to combinatorial optimization problem", Evolutionary Computation, 2000, Vol. 2, pp. 1354- 1360, 2000.

[5] Osamwonyi, I. O., & Evbayiro-Osagie, E. I. The Relationship between Macroeconomic Variables and Stock Market Index in Nigeria. J Economics, 3(1), pp. 55-63, 2012.

[6] Hui Song,Yingxiang Fan,Xiaoqiang Liu and Dao Tao. Extracting product features from online reviews for sentimental analysis, Computer Science and Converge Information Technology,2014, pp. 741-750.

[7] Phayung Meesad, Jiajia Li. Stock trend prediction relying on text mining and sentiment analysis with tweets, Fourth World congress on Information and Technology,2014,pp.257-262.

[8] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8 (2011).

[9] H. Alostad and H. Davulcu, “Directional prediction of stock prices using breaking news on twitter,” in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI- IAT)*, vol. 1, Dec 2015, pp. 523–530.