Indian Stock Market Prediction Using Machine Learning and Sentiment Analysis

Ashish Pathak and Nisha P. Shetty

Abstract Stock market is a very volatile in-deterministic system with vast number of factors influencing the direction of trend on varying scales and multiple layers. Efficient Market Hypothesis (EMH) states that the market is unbeatable. This makes predicting the uptrend or downtrend a very challenging task. This research aims to combine multiple existing techniques into a much more robust prediction model which can handle various scenarios in which investment can be beneficial. Existing techniques like sentiment analysis or neural network techniques can be too narrow in their approach and can lead to erroneous outcomes for varying scenarios. By combing both techniques, this prediction model can provide more accurate and flexible recommendations. Embedding technical indicators will guide the investor to minimize the risk and reap better returns.

Keywords Machine learning · Sentiment analysis · Stock market SVM

1 Introduction

This section describes the limitations of traditional approach in stock market analysis and lists the benefits of using machine learning and sentiment analysis.

1.1 Traditional Approach to Stock Market Analysis

Stock market is a very volatile in-deterministic system with vast number of factors influencing the direction of trend on varying scales and multiple layers. Efficient

A. Pathak (✉) · N. P. Shetty
Manipal Institute of Technology, Manipal University, Manipal 576104, India
e-mail: ashish.spathak33@gmail.com

N. P. Shetty
e-mail: pnishashetty@gmail.com

H. S. Behera et al. (eds.), *Computational Intelligence in Data Mining*,
Advances in Intelligent Systems and Computing 711,
https://doi.org/10.1007/978-981-10-8055-5_53

Market Hypothesis (EMH) states that the market is self-correcting, i.e. current stock price reflects the most relevant cumulative price which is neither undervalued nor overvalued, and any new information is instantly depicted by the price change [1]. In layman's term, "The market is unbeatable," as you cannot gain any advantage over the market, but existing research proves otherwise. It is possible to predict the market trends by analysing the patterns of stock movement. Traditional approach applies the following models for this.

- Fundamental analysis
  This approach focuses mainly on a company's past performance and credibility. Performance measures like P/E ratios are utilized to filter stock which may incline towards a positive price surge. This approach is based on theory that profitable companies will continue to be so because of uptrend influenced by rewarding nature of the market.
- Technical analysis
  This approach is based on predicting the future prices by applying time series analysis on previous trends. Statistical techniques such as Bollinger Bands, simple moving averages are applied to predict the successive trends.

1.2 Modern Approach to Stock Market Analysis

Computer science provides us with cutting-edge tools for machine learning like SVM and EML which can analyse and perform knowledge discovery at large scales in short amount of time. Two approaches for prediction of stock market are proposed in this research.

- Qualitative Analysis
  News feeds regarding stock market highly affect the market trend and thus form a downhill movement in case of a negative news. Thus, the media/social network and stock market data are highly coupled and make the system more unpredictable. Existing research points out that in case of crisis, stocks mimic each other and lead to market crashes [1]. Nowadays, Twitter has come forth as the most reliable and fastest way of consuming media. With combined resources of news feed and Twitter feed, general population sentiment about a company can be highlighted. Text mining and sentiment analysis are useful tools for such a high-scale analysis.
- Quantitative Analysis
  Historical data is now readily available for most markets. Using this dataset, we can apply multiple machine learning models to give accurate results for future investments. These models can be trained for individual stocks with adjusted bias for most reflective features. These models can also be trained to work in different scenarios and overall market movement.

Traditional approach focuses on fundamental analysis and technical analysis to predict the market at a large scale which rarely translates to low-level individual Stock Prediction, but it can be clearly observed that individual stocks contribute to whole market movement rather than the other way around. Thus, focusing on individual stocks to predict market movement is a much more logical approach. With technology advancing at such a rapid pace and abundance of computing power, we can now easily strive towards a comprehensive system to accurately predict the market trend and reap beneficial financial returns. Existing research proves that modern approach outperforms traditional approach and can output the most accurate results [1].

2 Literature Survey

Mehak Usmani et al. in [2] proposed an intuitive idea of combining results from historical data, news and Twitter feed sentiment analysis. This dual approach predicts the stock market trend with high accuracy. It uses technical analyses like ARIMA and SMA to get an idea of the market trend. These models forecast the values based on proven mathematical models. This research considers other factors like depreciation and exchange rates. This research utilizes technical analysis for prediction which has been proven inferior to machine learning in terms of accuracy. Machine learning can handle noise and lack of information more efficiently. This approach has chances of inaccuracy for market scenarios not covered in training data.

The work proposed in [3] by Rodolfo C. Cavalcante et al. improves upon previously existing trading rules and produces results better than research proposed before. This research uses multiple proven market strategies to stimulate a real-time autonomous trader. This research focuses on short-term gains which is excellent for hands-off trading. Their model accumulates lot of revenue by trading in small time frames (minutes). Improvements can be made on choosing more features and making it more flexible.

In [4], Paul D. Yoo et al. investigated the success of machine learning models and event-driven models like sentiment analysis in predicting the stock market trends. It also illuminates the fact that macroeconomic conditions like international and political events affect market trends and need to be taken into consideration.

Alexander Porshnev et al. in [5] stated that addition of Twitter sentiment analysis does not add any valuable information to the prediction model and does not increase the accuracy. Thus, this research takes news feeds into consideration to add credibility to sentiment analysis.

The research was done by Dongning Rao et al. in [6], which provides great insight into proper implementation of sentiment analysis. They propose increasing the size of corpus (training data) with each test. This is done by adding non-polarizing words found in the test data not present in the corpus. The training data is refined by doing K-crossfold validation during each testing phase.

3 Methodology

The aim of this project is to build an application which outputs accurate recommendations in a quantifiable manner. For this purpose, three modules are implemented which are as follows:

- Machine learning module
- Sentiment analysis module
- Fuzzy logic module

These modules are integrated into a recommendation model in the following manner as shown in Fig. 1.

3.1 Machine Learning Module

The purpose of this module is to output Stock Prediction value. Stock Prediction value is the strength of difference in opening price and closing price. For this, we need to predict the closing price of the stock. This is achieved by applying machine learning on historical data of the stock. Research in [3] affirms that maximum number of features required to accurately predict a stock's closing price for a specific day are given as follows:

1. Opening price of prediction day
2. Lowest and highest prices of the prediction day
3. Simple moving average
4. Exponential moving average of opening and closing prices of the prediction day
5. Exponential moving average of lowest and highest prices of the prediction day

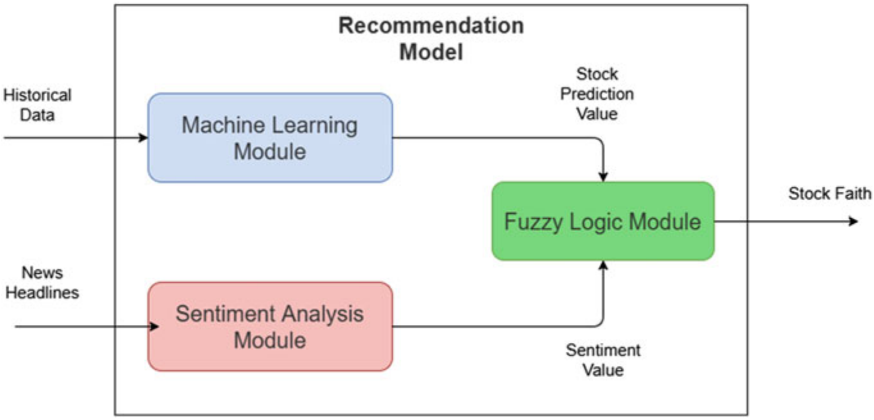

Fig. 1 Recommendation model for obtaining Stock Faith

6. Bollinger Bands of opening and closing prices of the prediction day
7. Bollinger Bands of lowest and highest prices of the prediction day

The training data is then fitted by a machine learning module and is used to predict the closing price of testing data through supervised learning. There are many regressors available **scikit-learn** library. Their accuracy was measured in terms of percentage error rate with accuracy calculated as shown in Eq. (1).

$$\left.\begin{array}{ll} \frac{|Predicted\ Closing\ Price - Actual\ Closing\ Price|}{Actual\ Closing\ Price} \times 100 < \alpha & Accurate \\ else & Inaccurate \end{array}\right\} \quad (1)$$

where α is the acceptable error rate.

On finding accuracy for α = 2 and 5, the accuracies observed are illustrated in Tables 1 and 2.

As it is obvious that Ridge Regressors give most accurate outcome for our dataset, it was selected to be used as the regressor for machine learning module to provide the Stock Prediction value.

The formula in Eq. (2) gives the Stock Prediction value.

$$\left(\frac{Actual\ Opening\ Price - Predicted\ Closing\ Price}{Actual\ Opening\ Price}\right) \times 100 + 50 \quad (2)$$

3.2 *Sentiment Analysis Module*

The purpose of this module is to obtain the sentiment value of latest news headlines regarding each stock and output its average as sentiment value to fuzzy module.

The steps used in this module are as follows:

1. Data Collection

The data is collected by crawling through Indian financial news Web site www.moneycontrol.com. Minimum four news headlines are scraped for each stock and stored against the company symbol.

Table 1 Accuracy table for closing price prediction (error rate less than 2%)

Classifier	Accuracy (%)
Lasso	40.79
LassoLars	51.61
Elastic Net	40.79
Ridge Regressor	85.4
SVR (kernel = linear)	0.97
SVR (kernel = RBF)	0.97
Random forest	15.44
AdaBoost	3.99
Decision Tree	3.67

Table 2 Accuracy table for closing price prediction (error rate less than 5%)

Classifier	Accuracy (%)
Lasso	64.03
LassoLars	72.49
Elastic Net	64.03
Ridge Regressor	94.2
SVR (kernel = linear)	2.37
SVR (kernel = RBF)	2.37
Random forest	29.49
AdaBoost	7.49
Decision Tree	9.29

2. Tokenizing
Each news headline is broken down into sentences and then in turn broken down into words.

3. Lemmatizing
It is the process of reducing inflected (or sometimes derived) words to their word stem, base, or root form. For example, "the boy's cars are different colours" reduces to "the boy car be differ colour."

4. Finding Most Informative Features
Words that contribute most in adding polarity to a sentence are found.

The top ten most informative features that contributed most to the polarity are listed in Table 3.

5. Classifying features into positive and negative
These are then classified into positive and negative using nltk packages.

6. Adding these features to the sentiment analyser lexicon
These words are then added to the sentiment analyser wordlist with appropriate strength for positive and negative words.

7. Classifying the testing data into positive and negative sentiments using training set
Now, our sentiment analyser is ready for classifying financial news from our sources.

Now to feed our sentiment value to fuzzy logic module, it needs to be normalized on a scale of 0–100 as shown in Eq. (3).

$$\left(\frac{\sum_{i=1}^{n}\left(Polarity(News_i)\right)}{n}\right)\times 100+50 \tag{3}$$

where n is the number of news articles pertaining to each stock.

Table 3 Most informative features

Positive	Negative
Buy	Sell
Up	Down
Rise	Dip
Jump	Hold
Strong	Bear
Support	Impact
Grow	Decline
Fold	Fall
Double	Loss
Bag	Debt

3.3 Fuzzy Logic Module

The purpose of this module is to output Stock Faith which is the strength of recommendation.

The activation rules for this module are as follows:

- IF the News Sentiment was good or the Stock Prediction value was good, THEN the Stock Faith will be high.
- IF the Stock Prediction value was average, THEN the Stock Faith will be medium.
- IF the News Sentiment was poor and the Stock Prediction value was poor, THEN the Stock Faith will be low.

Complete operation is illustrated in Fig. 2.

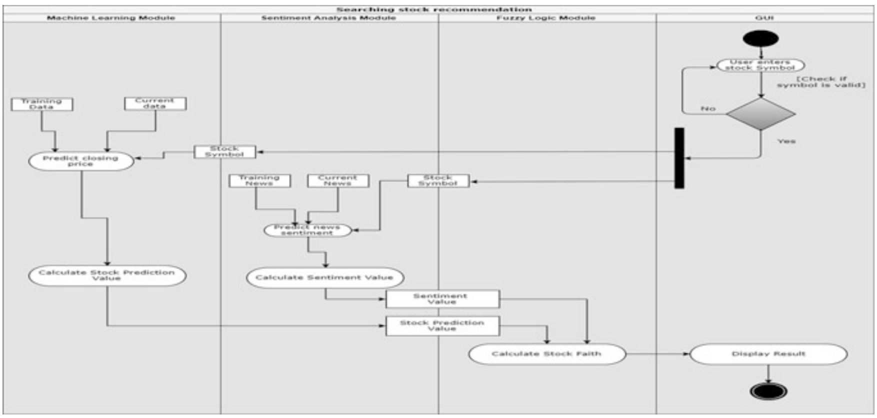

Fig. 2 Activity diagram

Symbol:	ABBOTINDIA		
opening price:	4225.0	closing price:	4267.65
high price:	4312.0	low price:	4225.0
predicted closing price:	4281.06		
sentiment Value:	85.0		
closing prediction:	51.33		
Stock faith:	62.47		
news1:	pos		
news2:	pos		
news3:	neu		
news4:	pos		
	BUY		

Fig. 3 Scenario for profit

Symbol:	BAJAJCORP		
opening price:	377.75	closing price:	373.2
high price:	377.75	low price:	371.5
predicted closing price:	373.07		
sentiment Value:	30.0		
closing prediction:	48.76		
Stock faith:	46.23		
news1:	neg		
news2:	neu		
news3:	neg		
news4:	neg		
	SELL		

Fig. 4 Scenario for loss

4 Result Analysis

Case 1: IF the News Sentiment was good or the Stock Prediction value was good, THEN the Stock Faith will be high as shown in Fig. 3.

Case 2: IF the News Sentiment was poor and the Stock Prediction value was poor, THEN the Stock Faith will be low as shown in Fig. 4.

5 Scope

National Stock Exchange of India (located in Mumbai) ranks at 12th largest in the world. NSE India has 1659 companies listed for public trading. Out of this, only 50 (known as Nifty 50) are focused on by investors. Nifty 50 acts as a barometer for Indian stock market growth. Indian economy relies mostly on exporting agricultural goods and services like software and technical support. Unfortunately, only 4% of India's GDP is derived from stock market exchange. This is much less compared to that of other developing countries which range from 20 to 40%. This untapped resource can be monetized more efficiently to contribute to the development of India.

6 Conclusion and Future Work

In this research, we propose that existing work [1–8] may integrate into a robust model to predict NSE stock market accurately. This model can be improved upon by defining refined fuzzy rules. Improving upon the training data's scale and time frame can result in better prediction. A trading model using the proposed methodology can be developed to compute total returns or investments in real time. This can prove the accuracy of the model. This model can successfully recommend the best stocks for investment.

References

1. Hellstrom, T. and Holmstrom, K. 1998, 'Predicting the Stock Market', Technical Report Series Ima-TOM-1997-07.
2. M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322–327.
3. R. C. Cavalcante and A. L. I. Oliveira, "An autonomous trader agent for the stock market based on online sequential extreme learning machine ensemble," 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, 2014, pp. 1424–1431.
4. P. D. Yoo, M. H. Kim and T. Jan, "Financial Forecasting: Advanced Machine Learning Techniques in Stock Market Analysis," 2005 Pakistan Section Multitopic Conference, Karachi, 2005, pp. 1–7.
5. A. Porshnev, I. Redkin and A. Shevchenko, "Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis," 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, 2013, pp. 440–444.
6. D. Rao, F. Deng, Z. Jiang and G. Zhao, "Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure," 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2015, pp. 381–384.
7. P. D. Yoo, M. H. Kim and T. Jan, "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation," International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vienna, 2005, pp. 835–841.
8. M. Qasem, R. Thulasiram and P. Thulasiram, "Twitter sentiment classification using machine learning techniques for stock markets," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 2015, pp. 834–840.