International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Predicting Indian Stock Prices through Sentiment Score and Data Mining Techniques

Govind S, Narayanan Prasanth, Sureshkumar WI, Balamurugan R

Abstract: ***Much is being said about the use of artificial intelligence to assist in the financial markets but there is a surprising lack of actual data supporting one trading algorithm over another. While there are numerous research that explore the possibilities of machine learning and deep learning techniques for stock market prediction, due to the lack of inter-domain expertise (people who are experts in both data science and finance), most of these projects use relatively elementary methods and fail to account for many underlying constraints. This paper aims to create a stock market investment aide which predicts the price action of stock market instruments using neural networks that learn time series patterns in the historical price data, correlates the prediction with the sentiment score of the market for that day and thereby give the investor/trader a buy or sell signal. The model developed has been successful in predicting price action of the NIFTY50 index with an accuracy of 89 percent. The prediction aided with the correlation from the sentiment analysis can give the investor an added confidence while making investment decisions on the stock market.***

Index Terms: Stock market, statistical methods, natural language processing, deep learning.

I. INTRODUCTION

Stock market is a highly complex, multi-dimensional monstrosity of complexities and interdependencies. Being of non-linear nature, stock market research to improve the risk to reward ratio has become an important issue in the recent years. In the financial world, analysts spent countless hours of collecting, segmenting, analysing and attempting to quantify the vast sea of data in the form of qualitative as well as quantitative information about companies in an attempt to maximize profits and minimize risks. This process, when manually performed by humans is tedious even though it is absolutely necessary. As a result, analysts are turning to use artificial intelligence techniques to make it do this laborious task for them. Deep learning techniques includes neural networks, natural language processing which is the sub-field of AI that is focused on enabling computers to understand and process human languages has proven monumentally effective towards achieving artificial intelligence driven fundamental analysis of stocks[1].

The vast amount of qualitative and quantitative information involved in analyzing a company means that it is virtually impossible for an individual to go through all of them to make informed decisions regarding investments. As a result financial institutions and investment banks are starting to employ algorithmic methods to do the work. These algorithmic methods are out of reach for individuals. As a result, it is becoming increasingly difficult for an individual investors to make money on the financial markets. Only if a news monitoring and stock prediction system designed from the perspective of an individual investor can attenuate this difficulty [2]. In this paper, the well-known efficient natural language processing and deep learning based approaches to stock market prediction are studied. A self-sufficient stock market prediction model is proposed that uses historic time series data of stock market price action and provides prediction of future price action and further buttress this prediction using the crowd sentiment that may affect the price action.

II. RELATED WORKS

A. Technical Analysis using Machine Learning Techniques

Technical analysis depends on historical data to predict the price movements. These include price, volume, and open interest statistical charts. This type of analysis is contingent on the idea that all the factors that are likely to affect the stock market are already factored into the price action of the historical data. Time series forecasting is a prediction method in which the past data of the stock price action is used to create a prediction variable which is analyzed and in turn modeled to observe the patterns that may be present in the historic changes. These models can in turn be used to forecast or predict the future prices. Two of the most commonly used time series modeling approaches are linear and non-linear approaches. Moving averages, time series regression and exponential smoothening are the commonly used linear methods. ARIMA or Auto-Regressive Integrated Moving Average model is one of the most popular linear method [3]. Even though this model presumes a linear model, it is very flexible since it can be used to model many different types of time series. These include Moving Averages, Autoregressive as well as a combination of the both called ARMA series.

Manuscript published on November 30, 2019.
* Correspondence Author
S.Govind*, School of CSE, Vellore Institute of Technology, Vellore, India.
Narayanan Prasanth N, School of CSE, Vellore Institute of Technology, Vellore, India
WI Sureshkumar, School of CSE, Vellore Institute of Technology, Vellore, India
Balamurugan R, Dept. of CSE, Vellore Institute of Technology, Vellore, India

The effectiveness of linear models like ARIMA for stock market analysis is minimal as stock market returns are rarely perfectly linear. This is because the residual variance between the actual returns and the predicted returns are very high. Therefore, non-linear models like neural networks have to be examined for stock market prediction applications.

Last few years have seen many advancements in the application of neural network models for stock market price forecasting with a hope that the price action of the stocks can be extracted successfully. The main advantage of using Neural Network models is that these models are capable of discovering non-linear relationships that may exist within the data set. Deep learning models have been successfully implemented to predict the stock price action. The network is able to classify within a certain proximity close matches in new, unseen data by unique properties of the visual representation of the stock price action i.e the candlestick chart. Hiransha M et al in their paper, NSE Stock Market Prediction Using Deep-Learning Models [1] uses four types of deep learning architecture i.e. Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for predicting the stock price. These models are applied to data sets collected from highly traded stocks of companies from three sectors namely Automobile, Banking and IT sectors from NSE. Mean Absolute Percentage Error (MAPE) is calculated and ARIMA model, which is a linear model is used for comparing linear and non-linear models. From the experiments on Indian stocks like HCLTECH, Maruti and Axis Bank, MLP network was found to be successful in capturing the seasonal pattern to some extend. RNN was found to be almost successful in identifying the pattern whereas LSTM and CNN fail to capture change in system between the specific periods of time but catches the overall pattern. In the case of New York Stock Exchange stocks like Bank of America, MLP network was found to fail to identify the pattern in beginning but later on it almost captured the pattern. RNN also fails in beginning time period and later on it captured the pattern but on reaching the end prediction was found to be a little lagging compared to the actual values. LSTM failed to capture the pattern at the beginning while CNN almost captured the pattern compared to other three networks. From the result it is evident that Deep Learning models have been capable to capture the pattern to some extend and is far superior to linear models. Out of the non-linear models used, CNN has been found to be the most effective.

Manna Majumder et al in their paper, Forecasting of the Indian Stock Market Index using Artificial Neural Networks [2] uses a multi-layer artificial neural network model to predict price index values of S&P CNX Nifty50 Index. The input variables selected for this model are lagged observation of the time series being forecasted i.e. the closing prices of S&P CNX Nifty 50 Index. The criticality in selecting the input variables lies in carefully selecting the number of input variables and the lag between each of these variables. The network architecture comprises of the input layer which contains the input neurons, two hidden layers with hidden neurons in each of the hidden layers and the output layer. The hidden layer of the neural network is associated with capturing the data patterns and characteristics, and establishing a complex dynamic nonlinear relationship between the input and output variables. The relationship between the input and the output of a neuron is established by the transfer function of the layer. The transfer function is a step function or a sigmoid function which takes the weighted input n and produces the output Y. Based on the performance of the network model transfer function are finalized for the network model.

In this paper back-propagation algorithm is used for training. A back-propagation network uses a supervised method of learning for training. The training algorithms considered for the study are:

1. Gradient descent adaptive back-propagation
2. Gradient descent with momentum and adaptive learning back-propagation.

The designed network model is validated for a period of 4 years. From the results obtained from the experiments, the author claims that a neural network model with 10 input variables and 1 output variable (with a linear transfer function) as well as 5 hidden intermediate neurons (with a tan sigmoid transfer function) is the most optimal. CNN has also been implemented for short term stock market prediction by many. Convolutional Neural Networks which has recently revolutionized the field of computer vision has been used to predict the price action from a picture of the time series of past price fluctuations. Neural networks are capable of classifying very complex relationships between characteristics of image/text and its corresponding classification. These relationships exceed the comprehension power of our brains as they are very subtle. Ashwin Siripurapu et al., In their research paper Convolutional Networks for Stock Trading [4] proposes feeding many samples of the companies' state with its future price outcomes. If the data in these networks is predictive, the CNN will be able to classify the "bullish" or "bearish" patterns in the past and subsequently recognize new buy and sell signals in the future.

The inputs to the model used by the author is the candlestick graphs of the stock prices in 30 minutes timeframes. A horizon of 5 minutes past the end of the window of past prices has been used to predict the future price. The price of the asset is taken as the average of the high and low price during each minute. Log returns have been used instead of arithmetic returns for accuracy with the neural network model. A loss function has also been implemented for more accuracy (l_2).

The workflow adopted by the author is given below:

1. Importing the candlesticks and converting image features and log return response into HDF
2. Generate network architecture file
3. Tune the hyper parameters
4. Train the network using Caffe
5. Visualize the weights in the trained networks
6. Evaluate the performance of the system.

The author received underwhelming results for the CNN model for short term stock market analysis. This shows that the most effective model can vary between time frames. LSTM (long short term memory) is another neural network model that can effectively model both long term and short term stock price data. The feature of LSTM which enables it is its ability to store, forget, and then read information from the long-term state of the underlying dynamics. Guanting Chen et al. in their paper,

Application of Deep Learning to Algorithmic Trading [5] effectively models the future stock price action using LSTM neural network model. The input features used by the authors are classified into historical trading data of INTC stock (open, high, low, close), commonly used technical indicators that are derived from the historical data and the index value of the market that the stock is traded on. The research has been done in three steps. Firstly, the best model has been chosen by training the network and evaluating the performance on a dev set. Then a prediction is made on a test set using the selected model. Lastly, the accuracy as well as profitability of the trained set is compared with a linear model - locally weighted regression model (LWR).

Since LSTM is a Recurrent Neural Network (RNN), the model is trained via back-propagation through time. This essentially means that for each cell, a fixed number of previous cells are unrolled and then a forward feed as well as back- propagation is applied to the unrolled cells. The experiment has been trained and tested on TensorFlow. The architecture of the LSTM used is 5 layers with 200 neurons. These numbers ensure that the network is deep and wide and can introduce necessary regularization to avoid overfitting or under fitting and can thereby improve the predictive accuracy. The authors claim that the LSTM accomplishes an accuracy of upto 80.328% which is exceptional.

B. Fundamental Analysis using Natural Language Processing Techniques

Fundamental analysis involves in-depth analysis of the changes of stock prices in terms of exogenous macroeconomic variables. This analysis assumes that the share price of an equity stock (or any other instrument) depends on the intrinsic value of the stock and the return expected by the investors. Text classification is the process of predicting a predefined class label, given an example of text. A popular example of text classification is sentiment analysis where class labels represent the emotional tone of the source text such as “positive” or “negative”. Sentiment analysis has been successfully applied to predict if a certain stock or index is “bullish” (buy signal) or “bearish” (sell signal) to some extent. In this paper “Analysis of Stock Market using Text Mining and Natural Language Processing” [6] Sheikh Shaugat Abdullah et al, propose a new data processing framework that takes text from different sources as input where the source may be authentic or unauthentic in the context of Bangladesh markets. Official information like changes in outstanding shares, bonus or dividend, mergers or acquisitions etc. is collected only from authenticated sources like the stock exchange’s website or official company website. Other information which reveals sentiment about the stock like rumors or chitchat which may or may not affect the stock price is collected from social media like Facebook and twitter as well as blogs and forums.

In subsequent layers, advanced filtering is performed on the data that is recently parsed before deciding whether a specific information/source is relevant or not. The users of the system must provide a watchlist which is a list of stocks that they are interested in investing in. The keywords from these entries are used to create a system log which filters out information about irrelevant stocks. Data from other stocks that are in the same industry as the pertinent stocks are not filtered as these may contain relevant information about how the stocks of that specific industry may move and therefore be crucial to decision making. Therefore, lexical as well as semantic resources of each of the categories is maintained in the database. Finally, an open source natural language processing tool called Apache OpenNLP has been used for extracting the information which is used for further analysis and calculations. For the decision making process from the parsed data, the latest available information is compared with the existing database to analyze the impact of a specific news on the stock market. In case of data from unreliable sources, the fundamental factors of the given stock is also evaluated. If the fundamental factors of the stock in question supports the data from these sources, then a higher weight is assigned to this decision. If not, the data from that specific source is ignored. The idea discussed in this research is different compared to others because both reliable and rumor/ unauthentic sources are taken into consideration. Further, a combined information parsing technique has been developed for both patterned and scattered text data types.

Many systems with similar procedure with light tweaks have been experimented with appreciable success. An example is the method introduced by Dongning Rao et al in their paper “Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure” [7]. Firstly, a corpus is entered as input. Named Entity Recognition is performed for each sentence using prescribed dictionaries. This process is called syntax based entity resolution. The initial dictionary is updated with common knowledge from different sources like the Palgrave dictionary of economics. After this, a classifier matrix and a correlation matrix is learned (supervised learning) from the common knowledge in the dictionary. The classifier is further applied into the corpus to obtain the prediction results. These prediction results are compared to a gold standard which could be a well-accepted set of labels that are made manually. The dictionary is updated to account for the differences that arise from the comparison. This process is continued until the differences between the prediction and the gold standard is minimal, at which point the learning process is terminated. At the end, the correlation matrix is produced as the output and the updated dictionary as the learned results.

While this paper presents a unified procedure for stock market prediction, testing has been held out for the future. Another drawback is over-fitting which is not accounted for in the process. Similar techniques have been used in the Arabic language by Alshahrani Hasan A et al. in their paper, “Sentiment Analysis Based Fuzzy Decision Platform for the Saudi Stock Market” [8]. In this paper a labeled corpora as well as a lexicon are manually built for the specific domain of Saudi stock market in Arabic. Analysis of the lexicon and the corpora has been done using two methods:

1. A corpus-based approach and
2. A semantical and lexical based approach.

The classified documents has further been used as an input to a decision mechanism for the Saudi stock market that employs fuzzy logic. Two corpora and one lexicon has been constructed and labeled. The first corpus has been taken from a Saudi investment forum called Saudishares, and the second one, from Twitter. The lexicon used, has been manually built out of the two contingent corpora to ensure robustness, accuracy, and reliability.

A content harvester called OutWit Hub has been used by the author to collect almost 19,000 posts for the dataset. These posts have been run through customized macros and scrapers to navigate through all the pages and pick only the bodies of the posts. The preprocessing stage in the proposed system consist of the following steps: removing unwanted parts (such as urls), removing stop words, removing non-Arabic characters, and tokenization. The lexicon has been primarily divided into two parts- positive and negative. Each of these parts are further split into two sub-parts. One of which contains the most unique words that alone gives a very clear indication about the polarity of the document and another which contains less intensive words. The words in the second part is given a lower score than the first one since they may be affected by features like negation.

Support vector machine (SVM) has been employed to classify documents that were harvested from 2 corpus- the twitter corpus and saudishares corpus. SVM itself is provided with a manually annotated corpora with about 80% of the data set for training and the remaining 20 % for testing

In their study, the authors have given more importance to recall, precision, and f-score than accuracy since they it is likely that accuracy may not be the best metric while making decisions (in this cause, to buy, sell or hold). The authors propose the use of the theory of fuzzy sets as the decision making mechanism. This theory is applicable because the membership of an element x(in this case stock price action) to a set U (buy, sell or hold) is likely to be partial (a value between 0 and 1) than a strict decision that it belongs to specific class. The classified documents are therefore entered into a fuzzy decision making mechanism.

Several supervised machine learning algorithms have been employed by the author and trained over the data set to calculate the F-score, recall and precision. The best F-score claimed by the author from the experiments is 63 percent and the best recall, 64 percent for rule-based method which uses IF-THEN conditions to classify data. This process goes through three main stages to obtain these metrics 1. Rule creation stage,

2. Rule ranking measure stage and
3. Classification stage,

The rule-based approach has been found to be the best way to classify documents and therefore give signals for action in the stock market for Arabic language.

A Japanese language implementation has been developed by Ken Maruyama et al through their paper “Is Stock BBS Content Correlated with the Stock Market?—A Japanese Case” [9]. This paper examines the correlation between an internet stock discussion board, Stock BBS and the price action of stock market in the Japanese scenario using various NLP techniques. The procedure followed in the proposed system has the following steps:

1. Morphological analysis and noise removal.
2. Feature vector calculation.
3. SVR classification.

Morphological analysis and noise removal is performed to extract words. The sentences are first divided into morphemes with the help of a morphological analyzer. Next, the noise removal is performed to remove the words that are unsuitable or of less importance.

Calculation of feature vector: The feature vector of each message used by the author has 6989 dimensions, and each value is the importance of a word. The importance is defined as the function of the appearance frequency of each word and calculate it by TF*IDF. TF increases the importance of a word that appears many times in a message. IDF is a filter of popular words because it increases the importance of a word that appears only in a specific message

Classification by SVR: The author postulates making use of the feature in Yahoo BBS by which the poster of each message in Yahoo! BBS can select a sentiment from five alternatives: “strong buy”, “buy”, “hold”, “sell”, and “strong sell”. However, a lot of messages are posted without sentiments. The learning data is messages with disclosed sentiments. Input data is the feature vectors of the messages, and the output for each message is “strong buy”=1, “buy”=0.5, “hold”=0, “sell”=-0.5, and “strong sell”=-1. Machine learning using SVR is done for all messages of all companies for which we can extract feature vectors. We classify messages having an SVR output equal to or larger than 0.5 as “bullish”, equal or smaller than - 0.5 as “bearish”, and other (from -0.5 to 0.5) as “neither”. Analysis is performed only on bullish markets and analysis on bearish market also has to be performed for the determination of the agreement index’s effects.

Two methods for stock market prediction using text classification NLP techniques for the Chinese language context are worth noticing. The first one has been presented in the paper “Impacts of Internet Stock News on Stock Markets Based on Neural Networks” [10] by Xun Liang. The paper explores the relationship between stock news on the internet and the stock markets in the Chinese context using neural networks. This is done in two steps. Firstly, impact of ISN on stock returns is evaluated. After this, the relationship between significant increases in the volume of internet stock news and respective changes in the stock prices is evaluated.

1. The first section examines the impact that Internet stock news has on stock returns. The author proposed that the contributors to the stock news should provide 5 headers that are reader-invisible while uploading the news. The 5 headers proposed by the author are the duration of impact, intensity of impact, range of circulation, status of the stock market and status of the industry. These 5 headers can be used to evaluate the impact of the news by inserting into a 5-H-1 feedforward neural network.
2. In the second section, a mapping between the significant increases of Internet Stock news and its correlation with the changes in the stock market is examined. This is done by using neural networks to evaluate the relationship between the daily volume of the stock news and the stock prices. The stock news have been harvested from a wide array of websites including ragingbull.com,
3. stockmoney.com and yahoo.com. The results exhibit a correlation between the both. The author claims that the training errors during the experiment is less than 12 percent and that the testing errors, less than 28 percent.

The results in this paper are preliminary and more studies has be done. A refinement of this method is explained by Xun Liang et al. in their paper “Mining Stock News in Cyberworld Based on Natural Language Processing and Neural Networks” [11].

This paper contributes to stock market prediction similar to [9] but with some improvements. The paper has been done in the Chinese context and therefore deals with Chinese natural language processing similar to [10]. In the proposed system in this paper, procedure of interpreting the web news for stock market prediction has been divided into two definite phases -

1. Harvest the stock news from the cyber world.
2. Application of neural network NLP techniques to harvested data set.

For processing the text in the Chinese natural language from the news pieces, two Chinese dictionaries - Chinese Word and Phrase Dictionary and Chinese word association dictionary are used. In addition to this open C++ codes for NLP developed by Chinese Academy of science is used to process the Chinese texts in the stock news articles. Another dictionary-user dictionary which contains about 430 typical Chinese stock related key words along with the associated strength in the interval [0, 1] is also selected by the authors in advance to obtain buy, sell or hold signals. Similar to [9] a feed forward neural network is used. The main difference is that a three layer perceptron is used in this case. The data of the web stock news is aggregated before it is given as input to the neural network. The neural network is capable of learning the neural mapping without any prior knowledge. In case the network cannot learn the pattern associated to a certain stock, a new hidden neuron is added. The authors claim that results of this experiment enhanced the results in [9] by over 7 percent at 69%. But the disadvantage of this method is that the structure of the neural network is more than twice as large as the network in [10].

A lexicon based sentiment analysis has been proposed by Sahar Sohangir et al. in their paper "Financial Sentiment Lexicon Analysis" [12] presents a new method for sentiment analysis for stock market prediction. Here data from Stocktwits (a website that aggregates market analyses from Twitter and then further condenses them into a curated and focused stream of data) is used as a dataset for the sentiment analysis. Machine learning based and lexicon based sentiment analysis is performed and therefore the dataset retrieved from the Stocktwits website is labelled. This paper investigates the relationship between Bullish/bearish tweets and positive/negative polarity through two approaches.

1. Machine learning approach: In this approach: the authors use these messages and supervised machine learning methods to classify StockTwits users' messages into either Bullish or Bearish sentiment. Linear Support Vector Machine, Naive Bayes and Logistic Regression methods have been applied. The performance of each of the methods have been found to be pretty close with an accuracy of prediction close to 80 %, F-measure around 90% and area under the curve around 70%.

2. Lexicon based approach: Three different methods of lexicon based sentiment analysis has been performed- TextBlob, SentiWordNet and VADER. Of these methods TextBlob was found to be largely ineffective due to it labelling too many messages as neutral. SentiWordNet as well as VADER was proven to improve the accuracy, precision as well as area under the graph by as much as 9% compared to the machine learning approaches. Out of these two, VADER is the most effective.

The primary drawback while using machine learning models for stock market prediction is the process of training which is often very time consuming as well as expensive computationally. Therefore, lexicon-based approaches are favourable while dealing with tasks that involve multi-dimensional data sets since they do not require training of the data. Based on the results of the study, the author claims that VADER not only outperforms normal machine learning techniques, but also other lexicon based methods used for extracting sentimental data from social media like Stock-twits.

III. PROPOSED MODEL

Although many research papers explore natural language processing and machine learning techniques to predict the price action of the stock market, most focus on either fundamental analysis or technical analysis alone. Pure technical analysis alone is not enough to consistently be right about the future since factors outside the market, such as politics, fundamentals, headlines and regulations may override any market internal analysis. Similarly, depending upon fundamental factors alone will make the investor oblivious to price action fluctuations. Therefore a model that uses fundamental information to corroborate technical analysis can help the investor make informed decision. The proposed framework consists of three modules and is shown in fig. 1.

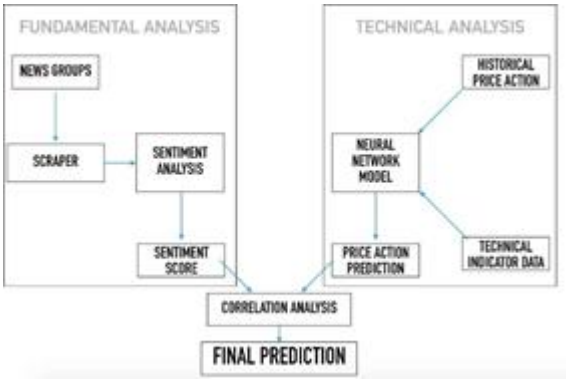

Fig. 1 Architectural diagram of proposed framework

A. Fundamental Analysis

The fundamental analysis module obtains financial news for any given day by scraping RSS feeds of reputed news websites. This is done with the help of feed parser and newspaper APIs in Python. The scraper is given a list of RSS feeds to scrape.

The list of financial RSS feeds to scrape is fed to the scraper using a .json file. The list contains the RSS feeds of many reputable financial news sites in India and is shown in fig. 2. The scrapers iterates through the list and scrapes a list of 10 news articles from each of the feeds which is shown in fig. 4. These scraped articles are further stored in another .json file in a dictionary format. This is done for ease of converting to a pandas data frame for sentiment analysis. The dictionary provides detailed information about each articles like title, content and date published.

For sentiment analysis, only the text data of each article is used. This is because, the titles can often be misleading. All other attributes of the articles in the dictionary are dropped and sentiment analysis is performed using python's VADER API. This API is a powerful lexicon based sentiment analyzer which uses semantic information from multiple corpora to analyze the sentiment of each sentiment. As a result, it does not require a training data set. This is importantto the model, as RSS feeds can be obtained only for the present day. VADER performs sentiment analysis of the given articles with respect to three lexicons- positive, negative and compound. From this, a positive, negative, neutral and compound sentiment scores are obtained for each article. Each of these sentiment scores are in the range -1 to 1. 1 being the most positive sentiment and -1 the most negative. The compound score is representative of both lexicons and the mean of the compound sentiment score for each article is averaged to obtain a sentiment score for the market for that day. A score between 0 and 1 implies that the market is likely to be bullish on that given day. A score between -1 and 0 implies that the financial world has a negative outlook on the market for that day and the price action is likely to be bullish. If the sentiment score obtained is 0, then there is neutral sentiment for the market on that given day. Table 1 shows the sentiment scores for first 5 articles scraped on 04-04-19 and Table 2 shows the date-indexed sentiment score of a particular article.

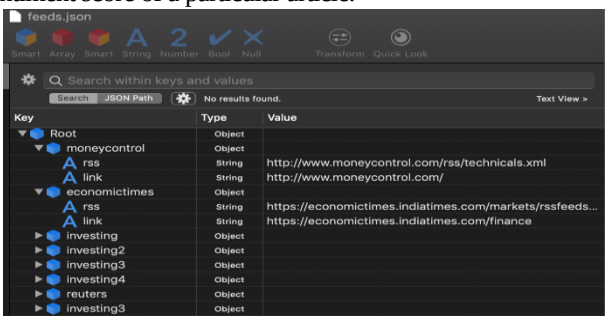

Fig. 2 RSS Feeds Of Financial News Websites That Are To Be Scraped

Fig. 3 Process of scrapper

B. Technical Analysis

The technical analysis module is concerned with predicting the price of a stock or any other financial instrument, by analysing the historic data and mining for patterns. The dataset used for the training, validation and testing is the historical prices of NIFTY50 index. This dataset is selected because it is the flagship index of the NSE of India. There are a number of ETFs that tracks the NIFTY50 and the predictions made by the model can be effectively used to invest in these ETFs for maximized earnings. The dataset is obtained from NSE's website. Because of the limitation of obtaining daily price movement for one year only at a time, yearly historic price data from 2008 until 2019 is obtained and collaborated.

Table 1 Sentiment scores for first 5 articles scraped on 04-04-19

	Compound	Negative	Neutral	Positive
0	0.8898	0.085	0.813	0.102
1	0.8067	0.057	0.863	0.080
2	0.9883	0.037	0.860	0.103
3	-0.0426	0.094	0.819	0.087
4	0.9955	0.052	0.790	0.158

Table 2 Date-indexed sentiment scores

Date	Comp	Neg	Neu	Pos
2019-02-05	0.9477	0.06	0.817	0.123
2019-02-06	0.9718	0.049	0.825	0.126
2019-02-07	0.7243	0.063	0.844	0.093
2019-02-08	-0.8934	0.1	0.842	0.058

The dataset has many irregularities like null values. Python's numpy API is used to preprocess the dataset by averaging the respective values of previous and next day's data. For now, the closing price is selected as the attribute on which the neural network trains tests and validates on. Rather than using closing price as is, the prices are normalized into percentage price changes for more effective learning and price prediction. The final dataset has 2804 days of closing price data which is divided into train-validate-test sets in the ratio 80:10:10. However, other distributions have to examine to see which is the most effective towards the prediction. After the training, the model predicts the price for the next day based on the information learned from the previous 200 days. This is made possible by the RNN as it has an internal memory unlike other types of neural networks like Feed-forward NN which is not capable of re-examining previous inputs. The parameters of the Recurrent Neural Network used is as given in table 3. As neural networks become increasingly accurate over time, this ensures that only the most recent learned data is used for predicting the future price action. After training and testing, the
model is back-tested against the validation data set which is 10% of the total dataset used. The price action prediction of the validation dataset is plotted for easy visual analysis as shown in fig. 4.

C. Correlation Analysis

The correlation analysis aims to correlate the outputs of the fundamental analysis module and technical analysis modules and further present the information in a form that is intuitive to the user.

This model is responsible for correlating the sentiment score obtained by the fundamental analysis model for the present day with the price predicted by the technical analysis module. This enables the model to provide a confidence level or risk level for each prediction which the user can take into account before making an investment.

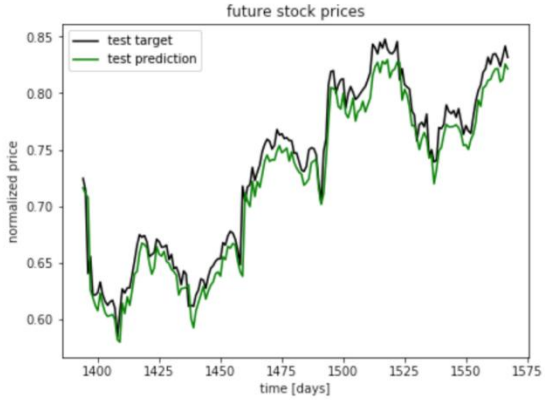

Fig. 4 Matplotlib plot of the prediction of future prices

Table 3 RNN Parameters

RNN parameters	Value
Sequence Length	20
No of steps	19
No of Inputs	4
Total number of neurons	200
No of outputs	4
No of hidden layers	2
Batch size	50
No of epochs for training	100

D. Assumptions and Constraints

D.1 Assumptions and Dependencies

This paper assumes that fundamental and technical analysis of data are both meaningful means to arrive at an investment decision. News feeds are assumed to be a fairly representative window towards the sentiment of the world regarding market direction or any financial instrument. It is assumed that historical data regarding the price action of the market or any instrument over a long period of time can provide valuable insight about the direction of the market or the instrument in the future. The accuracy of the prediction is highly dependent on the accuracy and validity of the news as well as historical data that is used to make the prediction. The more unstructured the data, the preprocessing requirements of the system increases. As preprocessing performed increases, there is a risk that the information may become incomplete and less representative of the actual situation.

D.2 Constraints

The availability of financial information from trusted sources is very limited with regards to the Indian financial markets. There is a complete lack of structured news information about individual stocks that can be effectively scraped from the internet in a form that can be used for analysis purposes. The most accurate historical charts and prices of any index, instruments or stocks that is available for free is from the website of the National Stock Exchange of India. The information from NSE is however, highly unstructured and requires extensive preprocessing before feeding into the model for learning purposes.

IV. RESULTS AND DISCUSSION

Sentiment Analysis has been performed on Financial RSS feeds using Vader. Sentiment API gives each article a compound, positive, negative and neutral sentiment score which is shown in fig. 5. The mean of these scores is taken to obtain a mean compound, negative, positive and neutral sentiment score. The compound score is representative of the overall crowd sentiment and this alone is used for final correlation analysis. The price action prediction using historic time series data and RNN gives a predicted percentage change of the closing price with respect to the opening price on any given day. The prediction is based on the information learned about the index over the last 200 days. As learning improves as time passes by, this ensures that the predictions are as accurate as possible. For individual stocks, the prediction of the price action could be done with an accuracy between 0.6 and 0.7 which is acceptable. Prediction accuracy for HDFC and Vedanta is shown in the fig.6 and fig.7 respectively.

```
x=np.mean(vs_compound)
print("Today's Compound Sentiment Score for the market is",x)

y=np.mean(vs_pos)
print("Today's Positive Sentiment Score for the market is",y)

z=np.mean(vs_neg)
print("Today's Negative Sentiment Score for the market is",z)

Today's Negative Sentiment Score for the market is 0.054028571428571426

w=np.mean(vs_neu)
print("Today's Neutral Sentiment Score for the market is",x)

Today's Neutral Sentiment Score for the market is 0.40111
```

Fig. 5 Sentiment Scores

But in the case of the NIFTY50 index, the prediction accuracy obtained is a very respectable 0.89. Each of the sentiment scores obtained for articles shows positive correlation with the predicted price action of the NIFTY50 index. But while predicting the price action for individual stocks, the positive correlation is low to nil. This is because, in the context of the Indian Financial Markets, Company specific RSS feeds are almost inexistent (except for Yahoo Finance). Table 4 shows the overall correlation of the predicted price action with sentiment score. This correlation can further be employed to recommend a strong or weak buy and a strong or weak sell.

If (close > 0 and sentiment > 0.1)
Print ("Strong Buy Signal. Closing price will most probably be higher than opening price")
dfx

Therefore, while using this algorithm, an individual can expect maximized return when their portfolios track every stock in the NIFTY50 index. While asset management firms can do this for investors, the benchmarks required is not accessible for individual investors.

Similarly, investing in a bag of stocks to track the index requires rebalancing to keep the asset weightage identical, this can be difficult to perform for individual investors, Last but not the least, NIFTY50 contains stocks with high market capitalization and individual stocks are considerable expensive. Therefore there are capital restrictions for new investors. Because of these reasons, the best investment decision that an individual investor can make when following this algorithm is to buy ETFs or electronically traded funds on the Stock Market. Since these instruments have low brokerage fees and low MER, it can reduce the transaction costs and therefore provide better net return each day. With the predicted price from the neural network model, correlated to the sentiment scores, we can obtain a. Although, the index price may not be representative of the price movement of individual stocks, it can be extensively applied to the trading of ETFs or Electronically Traded Funds. There are many funds that perfectly tracks the index values and insights about the price movements of index can be used to make decisions on when to buy and sell index tracking ETFs.

IV. CONCLUSION

Although the lack of structured financial news about individual stocks traded on the Indian financial markets makes it difficult to model the application for individual stock price prediction, index value price action such as of NIFTY 50's can be predicted accurately to some extent. The proposed model can help the investor crack into the comparatively new world of ETF instruments which are safer and free from the risk of institutional investor actions. Result shows the prediction accuracy reaches a maximum of 89%. However one of the limitation of the model is that historic RSS feeds are not archived for any financial news web sites in India. If these were available, the sentiment score for each day could have been appended to the dataset of the historic price charts for more accurate neural network modeling over 2 attributes. Future work must pertain primarily towards improving the data collection methods and sources, preprocessing and normalization methods as well as the use of different combinations of technical indicator information to provide a more accurate prediction.

REFERENCES

1. Hiransha M., Gopalakrishnan E., Vijay Krishna Menon, Soman K.P, "NSE Stock Market Prediction Using Deep-Learning Models", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), The NorthCap University, India.
2. Manna Majumder , MD Anwar Hussian, "Forecasting of Indian Stock Market Index using Artifical Neural Networks", NSE India, pp.1-21, 2010.
3. Asteriou, Dimitros; Hall, Stephen G., "ARIMA Models and the Box–Jenkins Methodology". Applied Econometrics (Second edition). Palgrave MacMillan. pp. 265–286, 2011.
4. Ashwin Siripurapu, "Convolutional Networks for Stock Trading", Stanford University Department of Computer Science, pp.1-6, 2017
5. Guanting Chen, Yatong Chen, and Takahiro Fushimi, "Application of Deep Learning to Algorithmic Trading, Fall Project Report, pp.1-6, Stanford University, 2017
6. Abdullah, Sheikh Shaugat, Mohammad Saiedur Rahaman, and Mohammad Saidur Rahman. "Analysis of stock market using text mining and natural language processing", International Conference on
7. Informatics Electronics and Vision (ICIEV), pp.1-6, May 2013.
8. Dongning Rao, Fudong Deng, Zhihua Jiang, Gansen 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, 2015
9. Alshahrani Hasan, Alvis C. Fong. "Sentiment Analysis Based Fuzzy Decision Platform for the Saudi Stock Market", IEEE International Conference on Electro/Information Technology (EIT), 2018
10. Ken Maruyama, Eiichi Umehara, Hirohiko Suwa, Toshizumi Ohta, "Is Stock BBS Content Correlated with the Stock Market?—A Japanese Case Ken Maruyama. "Is stock BBS content correlated with the stock market? — A Japanese case", IEEE International Conference on Systems Man and Cybernetics, Oct 2014
11. Xun Liang, Rong-Chang Chen. "Mining Stock News in Cyberworid Based on Natural Language Processing and Neural Networks", International Conference on Neural Networks and Brain, 13-15 Oct. 2005.
12. Ken Maruyama. "Is stock BBS content correlated with the stock market? — A Japanese case", 2009 IEEE International Conference on Systems Man and Cybernetics, Oct 2009
13. Sahar Sohangir, Nicholas Petty, Dingding Wang. "Financial Sentiment Lexicon Analysis", 2018 IEEE 12th International Conference on Semantic Computing (ICSC), 2018

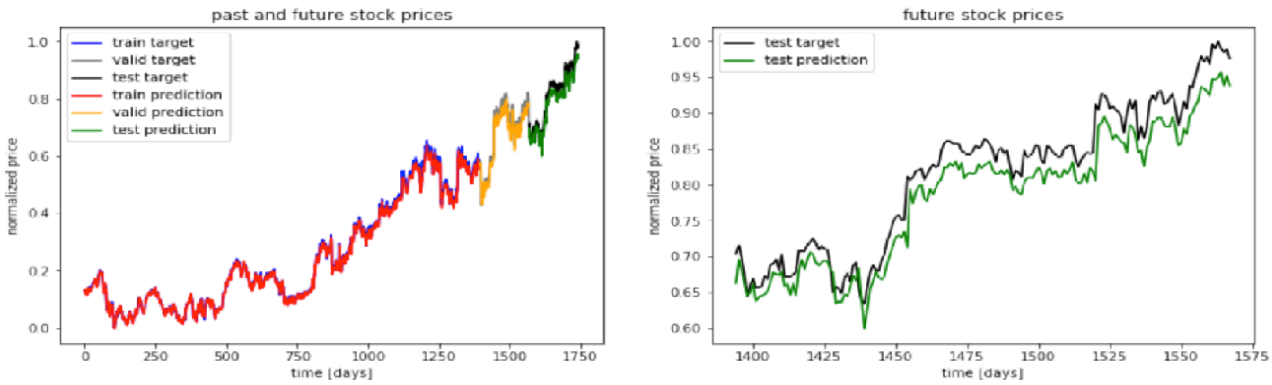

Fig. 6 Prediction Accuracy for HDFC

Fig. 7 Prediction Accuracy for NIFTY50

Table 4 Overall Correlation of the predicted price action with sentiment score

a. HDFCBANK

	PredictedScore	Comp
PredictedScore	1.000000	0.416118
Comp	0.416118	1.000000

b. NIFTY50

	PredictedScore	Comp
PredictedScore	1.000000	0.752547
Comp	0.752547	1.000000