Stock Market Forecasting Using Sentiment Analysis and Deep Learning

Veepin Kumar, Sanjay Singla, Shalika, Sandeep Kang, and Raman Chadha

Abstract Sentimental data processed from digital online communities can be used in different ways for market scrutiny. Sentiment Analysis is a way to extract opinion inclination (negative, neutral, positive) from a fragment of text cited for any institution or product. This word-of-mouth index can in turn be used to speculate the public mood and its market that has an impact on stock prices. Quarrying news articles and predicting the movements of product prices based on the content of Reviews corpus becomes beneficial. This research paper used Sentiment Analysis on the most popular Dow Jones Industries news articles to take advantage of this fact. We then combined this information with market index data from the company during the same time period to create a combined model that offers more accurate results based on Sentiment Analysis and Neural Networks (NN) and will assist the stockholder to lower risk and receive better returns.

Keywords Prediction · Neural network · Sentiment analysis and stock market

1 Introduction

The stock market means the collection of exchanges and markets where well-ordered activities of selling, buying of shares that are publicly held takes place along with their allocation. Stock markets offer a safe and highly structured environment where those who participate in the market negotiate and execute transactions in shares and

V. Kumar (✉)
Department of Information Technology, KIET Group of Institutions, Uttar Pradesh, Ghaziabad, India
e-mail: veepin.kumar@kiet.edu

S. Singla · S. Kang · R. Chadha
Department of Computer Science and Engineering, Chandigarh University, Chandigarh, Punjab, India

Shalika
Department of Computer Applications, KIET Group of Institutions, Ghaziabad, Uttar Pradesh, India

V. Goar et al. (eds.), *Advances in Information Communication Technology and Computing*, Lecture Notes in Networks and Systems 628,
https://doi.org/10.1007/978-981-19-9888-1_48

other eligible financial means with credibility with minimal execution menace [1]. The objective of stock market prediction is to determine the movement of stock value in the future of any financial exchange. Its accuracy in prediction can lead to greater profit for the investors. One of the most challenging issues is to predict the stock market trend due to various factors involved such as interest rates, economic growth, and politics that makes the market volatile and extremely hard to predict accurately [2].

In the modern economy, stock market information analysis and forecasting are crucial. Linear (ARIMA and MA) and non-linear models (ARCH and NN) can be used to categorize the many forecasting methods and algorithms [3]. There can be two major divisions of categories as to how the stock market can be predicted, these are namely technical and fundamental analysis. The first category of technical division aims at analyzing historical data of the stock prices for the prediction of future values, whereas the fundamental analysis depends on the examination of a collection of unstructured textual data, such as earning reports and financial news [4]. It is believed that technical analysis can predict the movement of the market but these research did not get high prediction results as they had a huge dependency on structured historical data not considering a very important source of information that consists of social media sentiments and online financial news [5]. Nowadays a large amount of critical information is available on the web related to the stock market including BBC, Yahoo Finance and Bloomberg. It gets hard to manually extract useful information from these sources making text mining for the same significant to extract useful information for analysis [6]. Modern approach to stock market Analysis inculcates both quantitative analyses using historical record of the past behaviour of the organization or product as well as qualitative analysis making use of news feed highly affecting the market trend with the growing influence of social media on prevalent scenarios. Predicting the future of the market comprises a vast range of variables including historic models to psychic models. But as stated by the behavioural economic hypothesis, market stats are related to public mood [3]. Social networking websites can provide a sizable collection of user generated content that in turn can be used for sample aggregation of public opinion for predicting market behaviour. Simple statistical forecasting using historical data based on a company's past performance ignores the importance of customer opinion (social media, news etc.).

In this paper we perform mood detection using lexicon-based sentiment analysis of the scrapped top news articles of the Dow Jones Industries and incorporate that intensity in a NN model that uses historical data to couple qualitative as well as quantitative analysis for our prediction of the stock market opening price for the next two months [7]. The forthcoming sections of this paper include the background of stock market and deep learning in Sect. 2, Sect. 3 explains the formulation of the problem and the suggested methodology. Section 4 describe the result analysis of our proposed solution and its effectiveness and at last, Sect. 5 concludes the research work and provide future directions.

2 Background

The study and analysis of literature is very essential since different researchers have utilised various datasets and approaches to forecast the stock movement trend through sentiment analysis. Huang et al. [1] test the Support Vector Machine's (SVM) capacity to predict the direction of financial movement by predicting the weekly movement direction of the NIKKEI 225 index and compare SVM's performance to that of linear discriminant analysis, quadratic discriminant analysis, and Elman backpropagation NNs in order to assess its predicting ability. The test outcomes show that SVM outperforms the other categorization methods. Khaidem et al. [2] made trials with a framework which anticipate whether stock prices will rise or fall with respect to the price prevailing in the past. Random forest algorithms and gradient boosted decision trees were used to produce results achieving an accuracy of 64%. In order to predict the closing price of an organisation based on previous prices, Hiransha et al. [3] used a number of deep learning models, including Recurrent Neural Networks (RNN), Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). CNN outperformed out of all the methods applied in this paper. The results were compared to the linear Autoregressive Integrated Moving Average (ARIMA) model, and it was shown that the NNs performed better than ARIMA. Pathak et al. [4] effectively performs a merge of quantitative as well as qualitative analysis of Indian stock market trends using sentiment analysis and machine learning (ML) with a fuzzy logic module. Sun et al. [6] look into the possibility of forecasting the stock market using text from user-generated microblogs. The model employed in the author's study differs from models used in past studies in two key ways: (1) it employs market information obtained in high-volume social media data rather than news stories, and (2) it does not analyse sentiment. The bulk of the companies listed in the S&P 500 index's data from 2011 to 2015 were used to evaluate the model, and it outperformed a baseline regression.

Usmani and Adil [8] instinctive design combines sentiment analysis results from historical data, news feeds, and twitter feeds. With remarkable accuracy, this strategy forecasts the direction of the stock market. To determine the direction of the market, it employs statistical time series models like ARIMA and SMA. In order to predict stock market trends, Yoo et al. [9] investigated ML models and event responsive algorithms like sentiment analysis. It also emphasises how factors like global and political events have an impact on market development and should be properly consideration. By analysing the text content of daily Twitter feeds using Opinion Finder, Bollen et al. [10] investigated the Dow Jones Industrial Average (DJIA) trend using emotional states from Twitter feeds. In order to investigate the prediction of changes in DJIA closing values by proposition of the public mood conditions measured by the Opinion Finder, a self-organizing fuzzy neural network is used. This network has an accuracy of 77.6 percent in predicting the daily up and down changes in the DJIA closing values. Table 1 list some of the research papers along with the technology used.

According to Porshnev et al. [13], the inclusion of twitter sentiment analysis does not improve the prediction model's accuracy and does not provide any useful

Table 1 Literature survey of different methods used for the problem

No	Author	Year	Technology used
1	Adebiyi et al. [11]	2009	MLP
2	Bollen and Mao [10]	2011	Opinion Finder + Fuzzy NN
3	Smailović et al. [12]	2013	SVM
4	Porshnev et al. [13]	2013	SVM + Neural Networks
5	Kumar and Anand [14]	2014	ARIMA
6	Arias et al. [15]	2014	SVMs and Neural Networks
7	Usmani and Adil [16]	2016	ARIMA and SMA
8	Pagolu et al. [17]	2017	ML
9	Lin et al. [18]	2017	SARIMAX + MLP

information. As a result, this study uses news feeds rather than tweets to increase the validity of sentiment analysis. The study provides a lot of insight into how sentiment analysis should be used. They suggested expanding the collection (training data) when each test was conducted, making the training data more effective as tests were conducted more frequently. Lin et al. [18] use the MLP model taking the sentiment feature alone with last four quarters historical sales number as our input feature.

3 Problem Formulation and Research Methodology

Our problem focuses on improving the application of a combination of quantitative as well as qualitative analysis for market trend determination and prediction. Earlier, there have been applications where this approach has been followed with different methodology and techniques to determine the results, for instance using time series analysis using SARIMAX model or by using MLP technique. We will predict the upcoming opening prices of the Dow Jones industries for a couple of months while considering sentiment indexes calculated using sentiment Analysis in our time series model using an efficient technique different from what has been already applied in the field to produce greater accuracy and reduced root mean square error. The overall architecture of the suggested problem is depicted in Fig. 1.

In this research work, two datasets have been used for the application of final methodology. The first dataset consists of the top twenty-five news articles of the Dow Jones stock prices for each day for a span of 8 whole years. The other dataset consists of stock market data of Dow Jones Industries with attributes including opening price, closing price, high and low prices, adjusted close price and volume for each day over a span of eight years. The news articles data was cleaned by removing unnecessary symbols (@, #, $), spaces, tags and URLs and were further processed to be analysed for opinion mining using NLP techniques. We applied an approach to map the lexical

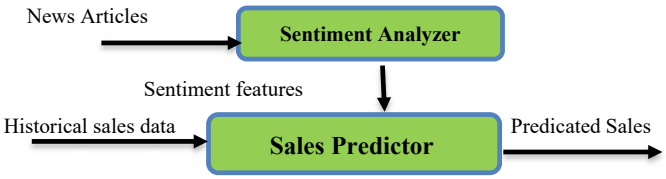

Fig. 1 Overall architecture of the problem

features to sentiment scores within a specific range constituting the range from a negative to a positive polarity of opinion. The key advantage of the lexical method is that since we already have everything, we don't need to train a model on labelled data; instead, all that would be left to do is evaluate the sentiment of sentences in the dictionary of emotions. The procedure for using lexical sentiment analysis is shown in Fig. 2. After processing both the datasets we merged the sentiment indexes and stock market data using matching dates. At the end of the pre-processing phase, the data frame contains attributes containing both stock price attributes and sentiment indexes for news articles computed for each matching day in the dataset. LSTM are well suited to classify, analyze, and predict time series with unpredictable time lags. Due to its relative gap length insensitivity, LSTM has an advantage over RNN, hidden Markov models, and other sequence learning techniques. Furthermore, LSTMs give us a variety of parameter choices, such as learning rates, input and output biases, and more. Consequently, there is no need for minor modifications. With LSTMs, updating each weight is now only O (1) in complexity.

A neural network stops learning due to the vanishing gradient problem because of which the updates to the various weights become smaller and smaller within a given neural network. This issue is mitigated with LSTM which makes it ideal for the purpose of predicting stock opening prices for forthcoming days using the stock values of previous records over a period. Problems with multiple input variables

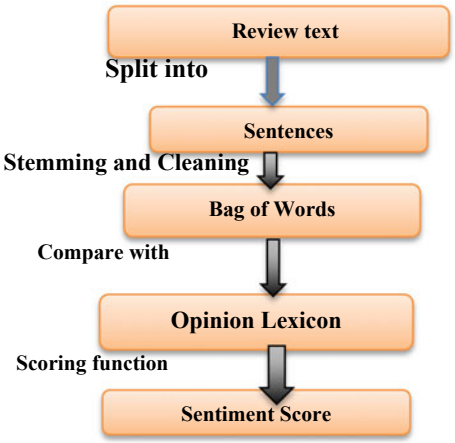

Fig. 2 Working of lexical sentiment analysis

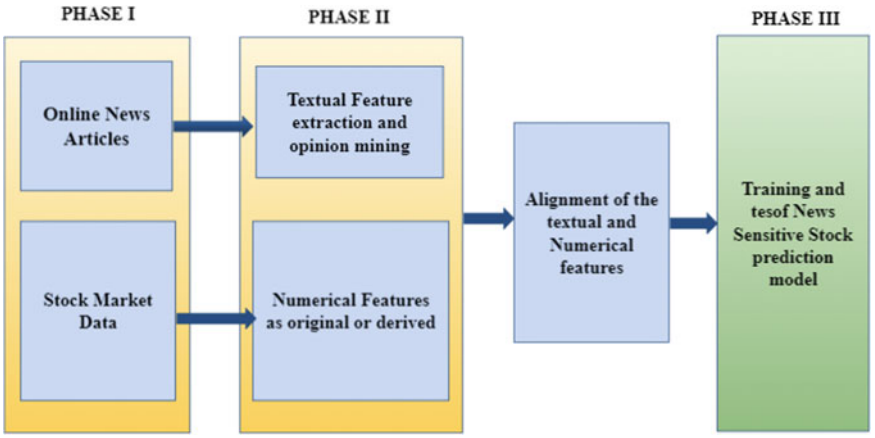

Fig. 3 Overall architecture of solution

can be coherently modelled using recurrent Neural Networks like LSTM. Traditional Linear methods can be hard to use with multiple input forecasting problems. Therefore, it can be a great benefit in time series forecasting. We use a multivariate LSTM Model and train it over different attributes governing Stock market movement to predict the faith of the market with its movement for the next two months. The solution of proposed problem is represented in Fig. 3.

4 Result Analysis

Following results were obtained after performing sentiment analysis on the collection of news articles from the dataset, categorizing them into positive, neutral, and negative as shown in Fig. 4.

The process of forecasting includes making future predictions based on recent and historical data. Networks are essential to assess patterns over time to comprehend the patterns in a long sequence of data. Recurrent networks, such as LSTM, which are used to learn such data, were employed. They can comprehend long-term interconnections and temporal differences. For our research, we trained the model on LSTM units over the span of 30 iterations using the Adam optimizer and Mean Squared Loss function.

Various algorithms have been used to implement a similar structure for stock forecasting and it was observed that the LSTM model provides a much lesser root mean square error than the other methods applied for the same. Figure 5 shows the prediction of a simple time series ARIMA model applied on the historical data of the stocks of Dow Jones Industries. The forecasted value has a large amount of root mean square error upon application rather than a closed curve with much lesser root mean square that is observed while applying the model proposed in Fig. 6.

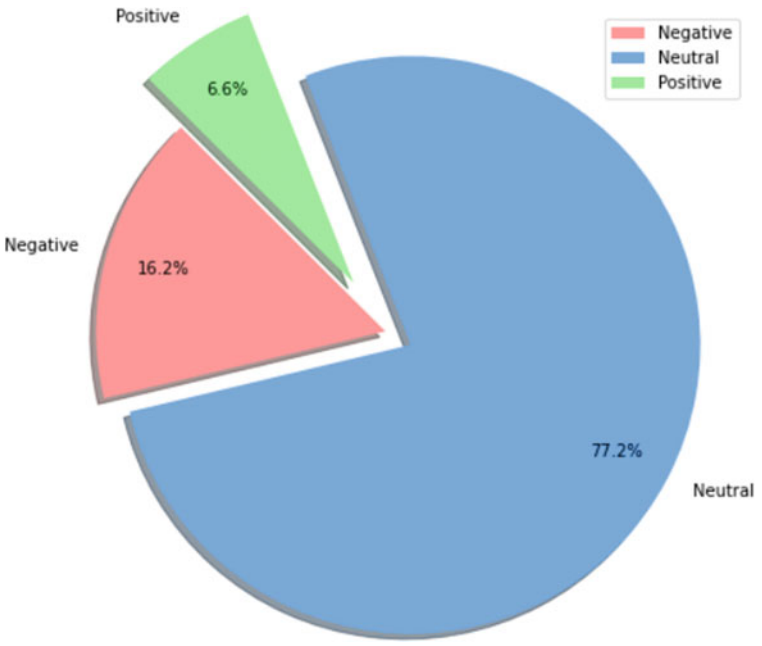

Fig. 4 Sentiment analysis of Dow Jones news headlines

Fig. 5 Training predictions versus actual stock prices with predicted stock price

5 Conclusion and Future Scope

Micro Blogging has become highly popular in today's era because it offers characteristic features like accessibility and satisfaction which allows instant dispersion of information and negligible response time with limited or no restrictions at all on content and security. Before the foundation of behavioural finance was established, research on stock market prediction was entirely based on unpredictable walks along

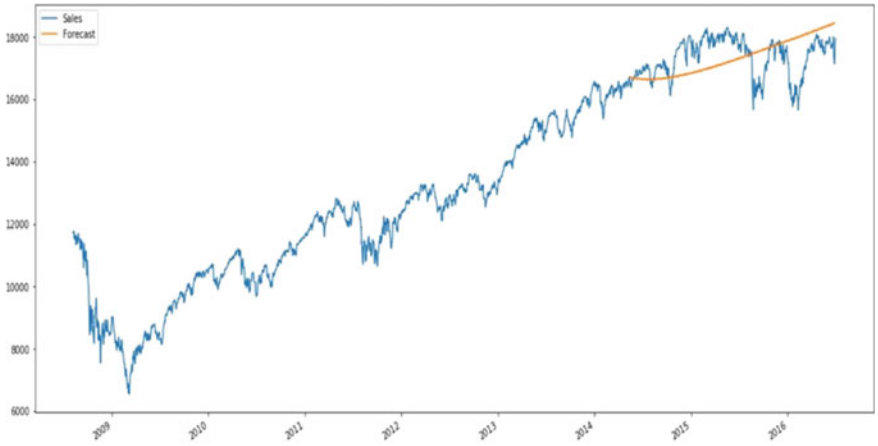

Fig. 6 Graph plot of predictions using ARIMA model

with numerical or quantitative prediction, but with its advent, people's beliefs that mood and emotions were also taken into account when predicting stock market movement. Therefore, we employed the concept of sentiment analysis of News Articles with Deep learning models to make it more effective. The results indicates that there is significant correlation linking the prevailing public emotions and etiquette of investment in the short term nevertheless this correlation is not proved to be significant statistically. Evidence also shows cause and effect relation between prevailing public sentiment and the stock market trends with respect to relationship between public emotions and daily closing prices. As part of a future implementation, we can perform comparative analysis with extreme learning classifiers which do not need gradient based backpropagation to function and deep learning classifiers using attribute pruning on the basis of variables that may be used for forecasting of market.

References

1. Huang W, Nakamori Y, Wang SY (2005) Forecasting stock market movement direction with support vector machine. Comput Oper Res 32(10):2513–2522. https://doi.org/10.1016/j.cor.2004.03.016
2. Khaidem L, Saha S, Dey SR (2016) Predicting the direction of stock market prices using random forest, 1–20 [Online]. Available: http://arxiv.org/abs/1605.00003
3. Hiransha M, Gopalakrishnan EA, Menon VK, Soman KP (2018) NSE stock market prediction using deep-learning models. Procedia Comput Sci 132:1351–1362. https://doi.org/10.1016/j.procs.2018.05.050
4. Pathak A, Shetty NP (2019) Indian stock market prediction using machine learning and sentiment analysis, vol 711. Springer, Singapore
5. Rao D, Deng F, Jiang Z, Zhao G (2015) Qualitative stock market predicting with common knowledge based nature language processing: A unified view and procedure. Proc—2015 7th Int Conf Intell Human-Machine Syst Cybern. IHMSC, vol 2, pp 381–384. https://doi.org/10.1109/IHMSC.2015.114

6. Sun A, Lachanski M, Fabozzi FJ (2016) Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. Int Rev Financ Anal 48:272–281. https://doi.org/10.1016/j.irfa.2016.10.009
7. Cavalcante RC, Ali O (2014) An autonomous trader agent for the stock market based on online sequential extreme learning machine ensemble. Proc Int Jt Conf Neural Networks, 1424–1431. https://doi.org/10.1109/IJCNN.2014.6889870
8. Usmani M, Adil SH (2016) Machine learning techniques using machine learning technique.
9. Yoo PD, Kim MH, Jan T (2005) Financial forecasting: advanced machine learning techniques in stock market analysis. 2005 Pakistan Sect Multitopic Conf INMIC. https://doi.org/10.1109/INMIC.2005.334420
10. Bollen J, Mao H (2011) Twitter mood as a stock market predictor. Computer (Long. Beach. Calif) 44(10):91–94. https://doi.org/10.1109/MC.2011.323
11. Adebiyi AA, Charles AK, Marion AO, Sunday OO (2012) Stock price prediction using neural network with hybridized market indicators. J Emerg Trends Comput Inf Sci 3(1):1–9
12. Smailović J, Grčar M, Lavrač N, Žnidaršič M (2013) Predictive sentiment analysis of tweets: a stock market application. Lecture Notes Computer Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7947. LNCS, 77–88. https://doi.org/10.1007/978-3-642-39146-0_8
13. Porshnev A, Redkin I, Shevchenko A (2013) Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. Proc—IEEE 13th Int Conf Data Min Work ICDMW 2013, pp 440–444. https://doi.org/10.1109/ICDMW.2013.111
14. Kumar M, Anand M (2015) An application of time series Arima forecasting model for predicting sugarcane production in India. Stud Bus Econ, 81–94
15. Arias M, Arratia A, Xuriguera R (2013) "Forecasting with twitter data. ACM Trans Intell Syst Technol 5(1). https://doi.org/10.1145/2542182.2542190
16. Usmani M, Adil SH (2021) Stock market prediction using machine learning. Procedia Comput Sci 194:173–179. https://doi.org/10.1016/j.procs.2021.10.071
17. Pagolu VS, Reddy KN, Panda G, Majhi B (2017) Sentiment analysis of Twitter data for predicting stock market movements. Int Conf Signal Process Commun Power Embed Syst SCOPES 2016—Proc, pp 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659
18. Lin Z, et al (2019) Learning to learn sales prediction with social media sentiment. Proc First Work Financ Technol Nat Lang Process, 47–53. Available: https://www.aclweb.org/anthology/W19-5508