Evolutionary Intelligence

Towards Ensemble Methods for Predicting One-Day Ahead Stock Movement using Google Trend Data

--Manuscript Draft--

Manuscript Number:	EVIN-D-18-00192
Full Title:	Towards Ensemble Methods for Predicting One-Day Ahead Stock Movement using Google Trend Data
Article Type:	Research Paper
Corresponding Author:	Satyendra Chouhan, PhD Shri Govindram Seksaria Institute of Technology and Science INDIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Shri Govindram Seksaria Institute of Technology and Science
Corresponding Author's Secondary Institution:	
First Author:	Aniruddh Nathani
First Author Secondary Information:	
Order of Authors:	Aniruddh Nathani
	Abhishek Gupta
	Avinash Saxena
	Satyendra Chouhan, PhD
	Jayendra Barua
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Predicting the direction of future stock prices is an important research topic in the area of trading, nance, statistics and computer science. Traditional stock prediction systems use time series data of a particular stock for predicting the movement of stock. However, these systems are less accurate. The recent research shows that the combining data from online sources (e.g., Google, wiki) related to a stock, can be eective for stock prediction. In this research, we analyze the stock prediction using machine learning techniques by combining traditional time series data, technical indicators and Google trends data. In addition, unlike existing research, proposed work presents an ensemble approach for predicting one day ahead movement of a stock. The experimental results show that (1) combining data from dierent sources is eective for predicting the stock movement; (2) ensemble approach is comparatively better then the individual machine learning models.
Suggested Reviewers:	

Towards Ensemble Methods for Predicting One-Day Ahead Stock Movement using Google Trend Data

Aniruddh Nathani[1], Abhishek Gupta[1], Avinash Saxena[1], Satyendra Singh Chouhan[1], and Jayendra Barua[2]

[1] Shri G.S.I.T.S Indore, 452001, INDIA
{avinashsaxena777,aniruddh.nathani1996,gupta2910abhi}@gmail.com,
schouhan@sgsits.ac.in
[2] Indian Institute of Technology (IIT), Roorkee, 247667, INDIA
jbaruadec@iitr.ac.in

Abstract. Predicting the direction of future stock prices is an important research topic in the area of trading, finance, statistics and computer science. Traditional stock prediction systems use time series data of a particular stock for predicting the movement of stock. However, these systems are less accurate. The recent research shows that the combining data from online sources (e.g., Google, wiki) related to a stock, can be effective for stock prediction. In this research, we analyze the stock prediction using machine learning techniques by combining traditional time series data, technical indicators and *Google trends* data. In addition, unlike existing research, proposed work presents an ensemble approach for predicting one day ahead movement of a stock. The experimental results show that (1) combining data from different sources is effective for predicting the stock movement; (2) ensemble approach is comparatively better then the individual machine learning models.

Keywords: Stock Prediction · Google Trends · Artificial Intelligence · Financial Expert System

1 Introduction

Stock prediction is an important research problem where the motivation is to predict the direction of future prices so that they can be bought and sold in order to make profit. Many experts normally use fundamental and/or technical analysis to predict stock movement and make investment related decisions.

Early researches such as Efficient Market Hypothesis (EMH) [7–9] showed that the stock prices cannot be determined as it follows a random path and the accuracy with which it can predict cannot exceed 50% [2].

However, state-of-the-art expert systems contradict the random walk theory and suggest that stock can be predicted effectively. Extracted by a time series analysis of their stock prices, these expert systems use technical indicators and the past prices of a stock [5, 14–16, 18, 20, 26].

In Addition, there are a lot of researchers who worked on the Knowledge based data from various data sources like Google, Yahoo, and Wikipedia etc. According to [22], we can predict the financial trading behaviour by observing Wikipedias financial related pages and its frequency of views, which was extended by [25] who observed the stock changing pattern by using the Google Trends. Most of the studies are performed on some particular stocks over a certain period of time. Moreover, they assume the subset of indicators form the available set of indicators.

Literature suggests that machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision tree (DT) are widely used for stock prediction and perform better than the traditional approaches [19, 10] and [27, 31].

In light of above work, we propose an ensemble based approach for predicting one day ahead stock price movement. We analyze the stock prediction using ensemble based machine learning techniques. The proposed approach combines the traditional time series data, technical indicators and features extracted from Google trends data. For the prediction, we use the two ensemble methods: Random Forest classifier and Adaboost classifier. We compare the performance of the proposed approach with some of the existing approaches.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the proposed work. Performance evaluation is given in Section 4. Section 5 presents the conclusion and future directions.

2 Related Work

Derivatives like options and functions, in modern finance, play a significant role not only in risk management activities but also in activities which involve price speculation. Although, predicting the financial markets movements is a difficult task, investors can gain huge profits with a small amount of input capital if they can precisely predict the market's direction owing to the high leverage characteristics involved in trading [30].

Academic investigations suggest that movements in stock market prices are not random. Moreover, studies show that they behave in a highly dynamic or a non-linear manner. According to a financial theory, the standard walk hypothesis, which states that stock market prices evolve according to a random walk i.e. price changes are random and thus impossible to predict. This hypothesis is merely a veil of randomness that obscures a noisy nonlinear process. To make the forecasting of futures prices more reliable and to remove this veil, the application of Artificial Intelligence (AI), such as expert systems, and neural network have received extensive attention.

Our focus, in this research, is on Machine Learning techniques for stock prediction. Thus hereafter we will discuss the state-of-the-arts that are closely related to the propose work.

There exists vast literature which focuses on stock market prediction. Many studies have used various types of ANN to precisely predict the direction of

movement of the stock prices. Moreover, predicting the stock price return, using ANN, has given us promising results.([1, 13, 24]).

Diler (2003) [6]estimated the direction of the ISE 100 Index using trained neural networks based on various technical indicators such as Moving Average Convergence-Divergence (MACD)RSI, Stochastics K%, MA, Momentum. The study presented results that the direction of the ISE 100 Index could be predicted at a rate of 60.81%.

Cao et al. (2005) [3] compared the capital asset pricing model (CAPM) and Fama and Frenchs 3-factor model to the predictive power of the uni-variate and multivariate neural network models. This was aimed at demonstrating the accuracy of ANN in stock price prediction for firms traded on the Shanghai Stock Exchange (SHSE). On comparison with linear models, the results presented that neural networks outperforms the linear models.

Huang et al.(2005) [12] aimed at evaluating the prediction ability of SVM. They investigated the predictability of financial movement direction with SVM by predicting the weekly movement direction of NIKKEI 225 Index. Then, they compared its performance with those of Elman back-propagation neural networks, linear discriminant analysis and quadratic discriminant analysis. The results of the experiment showed that SVM outperformed the other classification methods.

Kumar et al.(2006) [17] used random forest and SVM to predict the direction of the daily movement of S&P CNX NIFTY Market Index of the National Stock Exchange. Then they compared the results with those of the traditional discriminant, logistic models and ANN.

Hsu et al.(2009) [11]developed a two-stage architecture by integrating support vector regression and self-organizing map for stock price prediction. Seven major stock market indices were examined by them and the results showed that this two-staged architecture proves to be a promising alternative for stock price prediction.

Mittal (2012) and Si (2013) [21, 29] also used twitter data and showed that the social sentiment network using twitter data to find the one day ahead stock prediction and used non parametric topic based sentiment time series approach to analyse the streaming twitter data and used vector auto-regression model with learned twitter sentiments. However, there are very few works that combines the knowledge base obtained from different sources. Moreover, [32] also presented a prediction system that uses a knowledge from disparate sources for predicting one day ahead stock movement. It used machine learning model: ANN, SVM and DT for prediction.

Literature suggests that using difference source of knowledge base can be effective to predict the stock movement. Here we present the work that focuses on various knowledge bases. For example, there exist some works that suggests that early indication of stock movement can be extracted from online information (e.g., Google Trends and blogs). For e.g., Google search queries have been uesed to provide effective indicators of consumer spendings [4]. Schumaker et al. (2009) [27] shown that breaking news related to financial can be effective to

predict stock market movements. Moreover, Twitter feeds can be used to derive the mood related to the daily up and down in the DJIA[2].

In the light of above works, we present an ensemble based approach for stock movement prediction It combines the traditional time series data with feature extracted from Google trend data. The summarize comparison of proposed approach with state-of-the is shown Table 1.

Table 1. Comparison of propose work with state-of-the-art

	Knowledge Base		
Paper ID	Traditional	Web platform data (e.g., Google)	Approaches
Qian et al. (2007) [26]	Yes	No	ANN, DT
Li et al. (2008) [19]	No	No	ANN
Shynkevich et al. (2015) [28]	No	Yes	ANN SVM
Chourmouziadis et al. (2016) [5]	Yes	No	Fuzzy System
Nguyen et al. (2015) [23]	No	Yes	SVM
Bin Weng et al. (2017) [32]	Yes	Yes (News data)	ANN, SVM, DT
Proposed Approach	**Yes**	**Yes (Google Trend)**	**Ensemble**

3 Proposed work

The overall architecture of the propose technique is shown in Figure 1. The architecture consists of three phases; the first phase is data acquisition through different sources. The second phase calculates the features and technical indicators on the acquired data. In third phase, we apply two ensemble methods for stock prediction. The details of each phase is as follows:-

3.1 Data Acquisition

In this phase, we acquire data from two different sources; traditional time series data of a particular stock is obtain from the Alpha-Vantage website (*https://www.alphavantage.co/*) and Google trends data from Google. Thereafter features are calculated on traditional time series and Google trends data. The details are as follows:

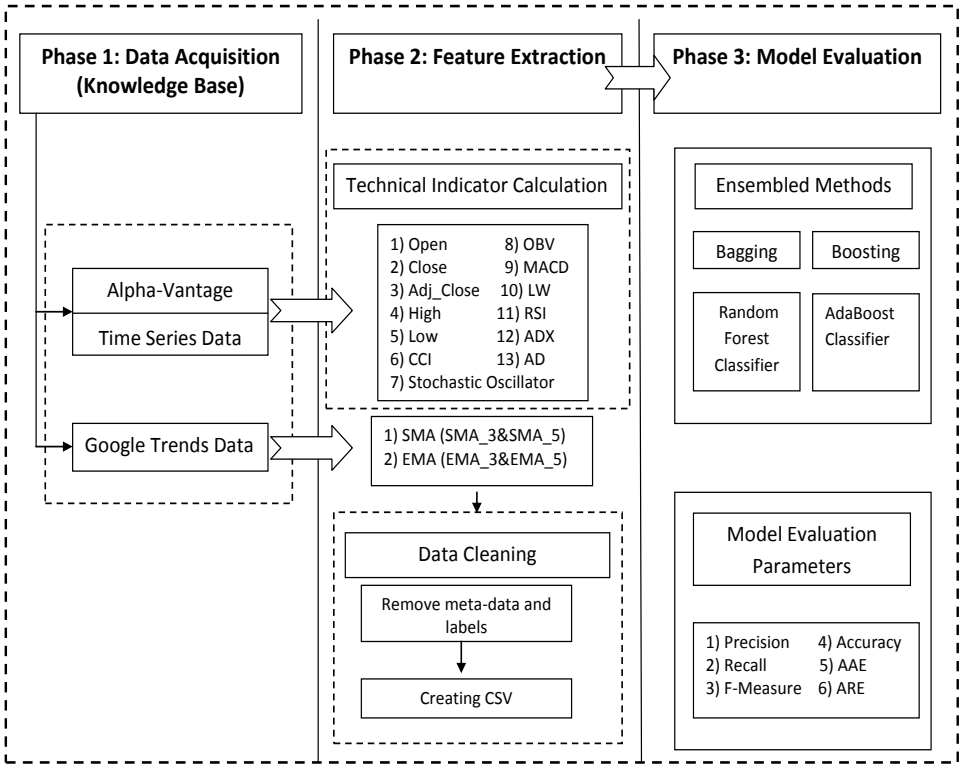

Fig. 1. An overview of Model Learning Process

Time Series Data One can define time series as a set of observations X_t , where each observation is recorded at a specific time t. Let the time series model for the observed data be $\{X_t\}$. It is thus a specification of the joint distributions of a sequence of random variables X_t. (X_t is a postulated realization) Also, it should be noted that the time series can only be observed at a finite number of times, and in that case the underlying random variables $(X_1, X_2, \ldots, X_n)$ is just an n-dimensional random variable. It is convenient to allow the number of observations to be infinite. In such a case $\{X_t, t = 1, 2, \ldots\}$ is called a stochastic process. In order to specify its statistical properties we then need to consider all the n-dimensional distributions-

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n \text{ for all } n = 1, 2, \ldots, n$$

There are several features that were obtained in the form of Time Series data format. We examine five different time series specifically-

- Open- Opening value of the stock for a particular day
- Close- Closing value of the stock for the particular day

- Adjusted Close - It makes use of the closing price as a starting point, and at the same time it takes into account factors such as dividends, stock splits and new stock offerings to discern its value.
- High- Highest value of the stock at that particular day
- Low- Lowest value of the stock at that particular day

Technical Indicators For future stock price forecasting, technical analysis is used. The second set of predictors consist of certain indicators that are used for technical analysis. This is done by analyzing previous prices and volumes.

Since stock prices provides all the necessary information, it is sufficient to study specific technical indicators (created using mathematical formula) to predict future price movements and evaluate the strength of the current trend. In this paper, we consider the following eight technical indicators:

1. Stochastic Oscillator (%K)- It is a momentum indicator, developed by George C. Lane, that can predict the strength or weakness of the market. When the market is moving upwards, it measures when the closing price would get close to the lowest price in a given interval. Whereas, when the market is moving downwards, it measures when the closing price would get close to the highest price for a given interval of time.

$$\%K = 100 \times \frac{close - LowestLow_{[last\ n\ periods]}}{HighestHigh_{[last\ n\ periods]} - LowestLow_{[last\ n\ periods]}}$$
$$\%\text{D} = \text{MovingAverage (\%K)}$$

2. Moving Average Convergence/Divergence (MACD)- It is a trend following momentum indicator that shows the relationship between two moving averages of prices. The Signal line is an Exponential Moving Average of the MACD which is calculated by subtracting the 26 day EMA. High values delineate overbought conditions whereas low values indicate oversold conditions. If the MACD is at extreme high or low values, such divergence with the price indicates an end to the current trend.
Formula:
$shortema = 0.15 \times price + 0.85 * shortema_{[-1]}$
$longema = 0.075price + 0.925 * longema_{[-1]}$
$MACD = shortema - longema$

3.The Larry William (LW) % R Indicator - This is a momentum indicator that provides for the identification of overbought and oversold levels. This indicator is commonly used to find entry and exit points in the market. Formula:

$$\%R = 100 \times \frac{HighestHigh_{[last\ n\ periods]} - close}{HighestHigh_{[last\ n\ periods]} - LowestLow_{[last\ n\ periods]}}$$

4. Relative Strength Index (RSI) - The RSI measures the velocity and magnitude of directional price and its value ranges from 0 to 100. The RSI indicator compares the magnitude of recent gains to recent losses so as to determine

overbought and oversold conditions of an asset over an allotted interval of time.
Formula:
If $close > close_{[-1]}$ then
$up = close - close_{[-1]}$
$dn = 0$
else
$up = 0$
$dn = close_{[} - 1] - close$
$upavg = \frac{upavg(n-1)+upn}{dnavg=dnavg(n-1)+dnn}$

$$RMI = 100 \times \frac{upavg}{upavg+dnavg}$$

5. Average Directional Movement Index (ADX)- Developed by Welles Wilder, ADX, Plus Directional Indicator (+DI) and Minus Directional Indicator (-DI),together, represent a group of directional movement indicators that form a trading system.
Formula:
$ADX = \frac{ADX_{[-1]} \times n-1+DX}{n}$

6.Commodity channel Index (CCI)- This indicator detects the beginning and the ending market trends. It ranges from 100 to -100 which is the normal trading range. CCI values outside of this range delineate overbought or oversold conditions.
Formula:
$CCI = \frac{TP-ATP}{0.015 \times MD}$
TP = highn + lown + close3
TP = Typical Price
lown = Lowest Low in the last n time periods
ATV = SimpleMovingAverage (TV)
MDTV = MeanDeviation (TV)

7.On Balance Volume (OBV)- It is a cumulative summation of the up and down volume. When the close is greater than the previous close, the volume is added to the running total, and when the close is lower than the previous close, the volume is subtracted from the running total.
Formula:
If $close > close[-1]$ then
OBV=OBV[-1] +Volume
elseif $close < close[-1]$ then
OBV=OBV[-1] -Volume
else
OBV=OBV[-1]

8. Accumulation/Distribution Line(AD)- Similar to the OBV, the AD Line indicator sums the volume times +1/-1 based on whether the close is greater

than the previous close. However, this indicator multiplies the volume by the close location value (CLV) which can be 0, +1 or -1 ans is based on the movement of the issue within a single bar.
Formula:
CLV = close - low - (high - close)(high - low)

AD = AD[-1] + (CLV x Volume)

Google Trends Data Google Trends is a trend analysis tool that shows how many times a specific search term or query is searched on Googles search engine in comparison to the sites total search volume over a given period of time.In other words it analyzes the popularity of a specific term searched on Google search across various regions and languages. We can also use it for comparative keyword research. It also provides query-related data and geographical information about search engine users. We used Google Trends to get the number of times a specific term related to company such as company name has been searched on Google.
Features calculated on Google trends data are-

Simple Moving Average (SMA)- The SMA indicator can be used efficiently to determine the direction of trend. A simple moving average is the arithmetic mean calculated by dividing the sum of recent closing prices by the number of time periods.
Formula:
$SMA = \frac{S1+S2+S3+S4+.....+Sn]}{n}$
where S1,S2.....Sn are google trends values and n is number of days.

Exponential Moving Average (EMA)- EMA or Exponential Moving Average gives a larger weightage to the recent observations as compared to SMA which gives an equal weightage to all the observations.
The three steps to calculating the EMA are:
1.Calculate the SMA.
2.Calculate the multiplier for weighting the EMA.
3.Calculate the current EMA.
Formula:
$EMA = [Closingprice - EMA(previousday)] \times multiplier + EMA(previousday)$

These indicators were calculated on the google trends data. For e.g., SMA3 calculated over data of recent 3 days is as shown below:-
SMA3_GoogleTrends The arithmetic mean of the closing prices of the most recent 3 days obtained via Google Trends data.
Formula:
$SMA3 = \frac{S1+S2+S3}{3}$

where S1, S2, S3 are the google trends value of the most recent 3 days.

EMA3_GoogleTrends The exponential moving average of the most recent 3 days, obtained via Google Trends data calculated by using the above formula for EMA.
The multiplier for EMA3 = 2(3 + 1)= 0.5

SMA5_GoogleTrends
The arithmetic mean of the closing prices of the most recent 5 days.
Formula:
$SMA5 = \frac{S1+S2+S3+S4+S5}{5}$
where S1, S2, S3, S4, S5 are the google trends value of the most recent 3 days.

EMA5_GoogleTrends
The exponential moving average of the most recent 5 days is calculated by using the above formula for EMA.
The multiplier for EMA5= 2(5 + 1)= 0.33

3.2 Ensemble classifiers

An Ensemble approach can be defined as a method built by combining multiple prediction techniques such that our prediction of the next day stock is more accurately determined. The basic idea of this approach can be related with the human nature looking for several options before making any decision. Ensemble uses the diversity of different prediction techniques to reduce the variance without increasing the bias of our prediction model.

The basic concept of this approach is that no single approach or system can claim to be superior to the another and thus the integration of several approaches will enhance the performance of stock prediction. In other words, this meta-algorithm combines several machine learning techniques into one predictive model and its main motives is to decrease variance (bagging), bias (boosting) or improve predictions (stacking). In our work we uses two ensemble classifiers.
1. Bagging (Using the Random-Forest Classifier)
2. Boosting (Using the AdaBoost Classifier

Bagging. (Bootstrap aggregating) was proposed by Leo Breiman in 1994 with an aim to improve classification by combining classifications of randomly generated training sets. It is a machine learning ensemble meta-algorithm which aims to manipulate the training data by randomly replacing the original T training data by N items. It involves generating different learning models from the different training subsets of given dataset. The final prediction is made by aggregating the decisions of all generated learning models. Each learning model is trained over a bootstrapped training subset. Following are the steps followed during bagging:

- Generating multiple training subsets from the available training dataset with the help of sampling with replacement.

- Training a different component of base learning technique using each subset.
- Combining the decisions of different learning models and predicting the final result.

Random Forest Classifier Developed by Adele Cutler and Leo Breiman, this method combines Leos bagging idea and the random selection of features, introduced by Tin Kan Ho, who initially proposed Random Decision forest in the year 1995. Random forests are an ensemble learning method for classification (and regression) that operates by constructing an assembly of decision trees at training time. Following are the two stages in Random Forest algorithm-

- Random forest creation,
- Making a prediction from the random forest classifier created in the first stage.

Steps for random-forest classifier
1. Random selection of X features from total Z features where X << Z
2. Calculation of the node d using the best split point, among the X features,
3. Splitting the node into daughter nodes using the best split.
4. Repeat the 1 to 3 steps until I number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for n number times to create n number of trees.

Steps for Random Forest prediction pseudo code:
1.Using the rules of each randomly created decision tree to predict the outcome by taking the test features and store the predicted outcome (target).
2.Calculation of votes for each predicted target.
3.Considering the high voted predicted target as the final prediction from the random forest algorithm.

Boosting. Another ensemble method which was proposed by Freud is Boosting. Prediction model built using this method involves sequential addition of one learning model at a time and using different example weights every time. It initially builds a learning model for a given training dataset and then subsequently the weights of the incorrectly predicted examples are increased in the next iterations.The weights are updated in such a manner that the error of incorrectly predicted examples is emphasized and counted in the coming iterations. Following are the steps involved in Boosting:

- Training a learning algorithm with training examples currently in hand.
- Compute the error of incorrect prediction and update the weights of training examples based on computed error rate.
- This process is repeated multiple times and final prediction is made by aggregating the decisions of the different learning models.

AdaBoost Classifier. Also known as the Adaptive Boost, this classifier is a type of boosting algorithm and is generally used with short decision trees. The first tree is created and the performance of the tree in each training instance is

used. Also, it is used to decide how much weight should be assigned to the next tree. Harder it is to predict the training data, more is the weight given to it. Alternatively, easy to predict instances are given less weights.

Models which are weak, are trained using the weighted training data and are added sequentially. This process continues until a pre-set number of weak learners have been created (a user parameter) or no further improvement can be made on the training data-set. Finally, there exists a pool of learners at the end, each with a stage value.

4 Experiments and Results

In this section, first we present the experimental setup and data set description. Next, we present the experimental results of two ensemble models: Random Forest classifier and ADA boost classifier. Thereafter we compare the best performer among the ensemble with individual classifiers.

4.1 Experimental setup

We have used Python programming language for building and evaluating different ensemble methods and comparing them with other learning techniques and algorithms like Knn, Logistic regression and Decision Trees. The experiments are conducted on a machine with Intel Core i7 CPU 3.4 GHtz, Windows 10 platform. We have used the AXIS BANK LTD., NSE: AXISBANK, stock movement based on a period of 10 years.

The data contains 2818 labeled stock dataset with past time series data over the period 10 years; along with the technical indicators computed on the time series data and features extracted on Google Trends data. We obtain publicly available market data on AXIS BANK using the Alpha-Vantage website for real-time and historical stock data. The details of the dataset is presented in Table 2. The full dataset is available at: https://github.com/AniruddhNathani/Stock_prediction_Using_Ensembled_Approach.

We have several one-day-ahead outcomes that can be of interest to several investors in the stock market. We have taken into account a single target. We compare the opening prices of two consecutive trading days. It is important to note that we have calculated this target only for AXIS BANK stock as a case study. Also, we have transformed this target to a binary variable where 1 refers to an increase in the target and 0 refers to no increase in the target value from the previous day.

For experimental purpose, we consider three scenarios that are presented in Table 3. In scenario 1, time series data used to obtain the performance for each of the ensemble models. In scenario 2, time series data and technical indicators are considered for the evaluation of learning models. While in scenario 3, time series data, technical indicators and feature extracted from Google trend data are considered for learning models. In each of the three scenarios, models are

Table 2. Dataset description

Time Series Features	**Technical Indicators**	**Google Trends Indicators**	**Target Variable**
Open,	Stochastic Oscillator,	Simple Moving Average(SMA) SMA_3SMA_5,	Open(i+1)-Open(i): where i and i+1 refer to two consecutive trading days.
Close,	Moving Average Convergence/Divergence (MACD),	Exponential Moving Average (EMA) EMA_3EMA_5,	
Adjusted Close,	The Larry William (LW) Indicator,		
High,	Relative Strength Index (RSI),		
Low,	Average Directional Movement Index (ADX),		
	Commodity channel Index (CCI),		
	Accumulation/Distribution Line(AD),		
	On Balance Volume (OBV),		

Table 3. Scenarios for performance evaluation

Scenario No.	**Description**
Scenario 1	Time Series Data of the stock.
Scenario 2	Time Series Data of the stock + Technical Indicators as additional features
Scenario 3	Time Series Data of the stock + Technical Indicators + Features calculated on the Google Trends data

evaluated using performance parameters such as : accuracy, average absolute error (AAE), average relative error (ARE), accuracy, precision, recall and f-measure. The description of the parameters are given in Table 4.

4.2 Results and Analysis

In this section, we present the experimental results of ensemble models on the three scenarios (Table 3).

Table 5 shows the results; from the results we observe that:

– For Scenario 1 we see that the AdaBoost Classifier was outperformed by the RandomForest Classifier. The 4 measures Accuracy, Precision,Recall and F1 score yielded a higher value for the RandomForest Classifier whereas Error results AAE and ARE were found to be higher in case of AdaBoost Classifier.

Table 4. Performance evaluation measures

Performance measure	**Description**
Accuracy	$\frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$ i.e., $\frac{TP+TN}{TP+TN+FP+FN}$
precision	$\frac{TP}{TP+FP}$
recall	$\frac{TP}{TP+FN}$
f1-score	$\frac{2*(recall*precision)}{(recall+precision)}$
AAE (Average Absolute error)	$\frac{\sum_{i=1}^{n}\lvert Y_i - X_i\rvert}{n}$ Where Yi represents the predicted value and Xi represents the actual values.
ARE (Average relative error)	$\frac{AAE}{Known\ values}$ i.e., $\frac{AAE}{Y_i+1}$

*TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

- For Scenario 2 we see that both AdaBoost Classifier and the RandomForest Classifier performed equally Well. Both the Classifiers yielded comparable values for all the 6 measures calculated on our dataset.
- Finally, for Scenario 3 we observe that that our AdaBoost Classifier outperformed the RandomForest Classifier. The 4 measures Accuracy, Precision,Recall and F1 score yielded a higher value for the AdaBoost Classifier whereas Error results AAE and ARE were found to be higher in case of RandomForest Classifier.

Table 5. Experimental results

	Scenario 1		**Scenario 2**		**Scenario 3**	
	Random Forest	**AdaBoost**	**Random Forest**	**AdaBoost**	**Random Forest**	**AdaBoost**
Accuracy	0.721	0.523	0.716	0.718	0.733	**0.751**
Precision	0.723	0.521	0.727	0.72	0.732	**0.75**
Recall	0.724	0.523	0.728	0.723	0.73	**0.754**
F1 - Score	0.721	0.516	0.726	0.72	0.732	**0.752**
AAE	0.288	0.586	0.293	0.291	0.276	**0.259**
ARE	0.199	0.284	0.200	0.206	0.202	**0.188**

Also, from the results of the three scenarios in (TABLE no.) we see that the Scenario 3 in which we have used Time-Series data along with Technical

Indicators and Features calculated on the Google Trends data, yields a higher accuracy. In the other two scenarios in which we used only the Time Series data (Scenario 1) and combined Time Series data with Technical Indicators (Scenario 2), the accuracy of our prediction is comparatively low and error measures are comparatively high.

From this we can conclude that prediction using the features calculated on Google trends data, is clearly a better method to inform us about the stock movement of the next day.

Comparative analysis. Since Scenario 3 yielded higher results, we expanded our research on Scenario 3 and calculated the accuracy for our model using the traditional approaches and individual techniques such as Decision trees, Logistic regression and SVM. Table 6 shows the comparison results.

Table 6. Comparison with individual classifiers

	Decision Tree	**KNN**	**Logistic Regression**	**Bagging using RandomForest**	**Boosting using AdaBoost**
Accuracy	0.639	0.493	0.521	0.733	**0.751**
Precision	0.64	0.492	0.521	0.732	**0.75**
Recall	0.641	0.491	0.52	0.73	**0.754**
F1 - Score	0.64	0.49	0.521	0.732	**0.752**
AAE	0.365	0.507	0.479	0.276	**0.249**
ARE	0.268	0.380	0.344	0.202	**0.188**

From the results we observe that:

- The KNN (K Nearest Neighbors) and LR (Logistic Regression) Algorithms yielded comparative results. But both these algorithms performed very differently for our Scenario 3 as compared to other algorithms used by us. It yielded a very low accuracy and the error rates were very high.
- The DT algorithm , it performed significantly better as compared to KNN and LR Algorithms. But the error rates were still pretty high so there was a scope of improvement in the methodology.
- Finally, moving on to the Ensembled methods which was the main methodology we focussed on, we see that these methods (RandomForest and AdaBoost) outperformed KNN, LR and DT. We observe that there was a significant increase in the values of the parameters. Also, the error rates also dropped down by a great extent.

Thus the above comparative analysis proves ensemble learning techniques using Random Forest and AdaBoost have outperformed all these individual approaches and hence we conclude that ensemble learning is clearly a better approach. We could achieve an accuracy of about 75% using the Boosting approach and about 73% using the Bagging approach.

5 Conclusion

In this paper, we presented an ensemble based approach for stock movement prediction. The significance of the proposed approach is to use the combination of various knowledge base for prediction. We have evaluated the proposed approach on a particular stock data set. The experimental results show that: 1) ensemble approach works better than the other individual classifiers for the prediction of stock movement and 2) combining the Google trend data into traditional time series data improves the performance of the prediction system. The future prospects of our work can include data from various social media platforms (e.g., Facebook, Twitter, etc.) to improve the accuracy of the prediction.

References

1. Avcı, E.: Forecasting daily and sessional returns of the ise-100 index with neural network models (2007)
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of computational science **2**(1), 1–8 (2011)
3. Cao, Q., Leggio, K.B., Schniederjans, M.J.: A comparison between fama and french's model and artificial neural networks in predicting the chinese stock market. Computers & Operations Research **32**(10), 2499–2512 (2005)
4. Choi, H., Varian, H.: Predicting the present with google trends. Economic Record **88**, 2–9 (2012)
5. Chourmouziadis, K., Chatzoglou, P.D.: An intelligent short term stock trading fuzzy system for assisting investors in portfolio management. Expert Systems with Applications **43**, 298–311 (2016)
6. Diler, A.: Predicting direction of ise national-100 index with back propagation trained neural network. Journal of Istanbul Stock Exchange **7**(25-26), 65–81 (2003)
7. FAMA, E.: Efficient capital markets: Ii the journal of finance. Vol. XLVI (5), 1575–1611 (1991)
8. Fama, E.F.: The behavior of stock-market prices. The journal of Business **38**(1), 34–105 (1965)
9. Fama, E.F., Fisher, L., Jensen, M.C., Roll, R.: The adjustment of stock prices to new information. International economic review **10**(1), 1–21 (1969)
10. Guresen, E., Kayakutlu, G., Daim, T.U.: Using artificial neural network models in stock market index prediction. Expert Systems with Applications **38**(8), 10389–10397 (2011)
11. Hsu, S.H., Hsieh, J.P.A., Chih, T.C., Hsu, K.C.: A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. Expert Systems with Applications **36**(4), 7947–7951 (2009)

12. Huang, W., Nakamori, Y., Wang, S.Y.: Forecasting stock market movement direction with support vector machine. Computers & Operations Research **32**(10), 2513–2522 (2005)
13. KARAATLI, M., GÜNGÖR, İ., DEMİR, Y., KALAYCI, Ş.: Estimating stock market movements with neural network approach. Yönetim ve Ekonomi Araştırmaları Dergisi; Sayı: 3; 38-48 (2005)
14. Kim, K.j.: Financial time series forecasting using support vector machines. Neurocomputing **55**(1-2), 307–319 (2003)
15. Kim, K.j., Han, I.: Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert systems with Applications **19**(2), 125–132 (2000)
16. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural networks. In: Neural Networks, 1990., 1990 IJCNN International Joint Conference on. pp. 1–6. IEEE (1990)
17. Kumar, M., Thenmozhi, M.: Forecasting stock index movement: A comparison of support vector machines and random forest (2006)
18. Lee, K., Jo, G.: Expert system for predicting stock market timing using a candlestick chart. Expert systems with applications **16**(4), 357–364 (1999)
19. Li, S.T., Kuo, S.C.: Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based som networks. Expert Systems with applications **34**(2), 935–951 (2008)
20. Lin, X., Yang, Z., Song, Y.: Intelligent stock trading system based on improved technical analysis and echo state network. Expert systems with Applications **38**(9), 11347–11354 (2011)
21. Mittal, A., Goel, A.: Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf) **15** (2012)
22. Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T.: Quantifying wikipedia usage patterns before stock market moves. Scientific reports **3**, 1801 (2013)
23. Nguyen, T.H., Shirai, K., Velcin, J.: Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications **42**(24), 9603–9611 (2015)
24. Olson, D., Mossman, C.: Neural network forecasts of canadian stock returns using accounting ratios. International Journal of Forecasting **19**(3), 453–465 (2003)
25. Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using google trends. Scientific reports **3**, 1684 (2013)
26. Qian, B., Rasheed, K.: Stock market prediction with multiple classifiers. Applied Intelligence **26**(1), 25–33 (2007)
27. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Transactions on Information Systems (TOIS) **27**(2), 12 (2009)
28. Shynkevich, Y., McGinnity, T.M., Coleman, S., Belatreche, A.: Stock price prediction based on stock-specific and sub-industry-specific news articles. In: Neural networks (ijcnn), 2015 international joint conference on. pp. 1–8. IEEE (2015)
29. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., Deng, X.: Exploiting topic based twitter sentiment for stock prediction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 24–29 (2013)
30. Tsaih, R., Hsu, Y., Lai, C.C.: Forecasting s&p 500 stock index futures with a hybrid ai system. Decision Support Systems **23**(2), 161–174 (1998)

31. Vu, T.T., Chang, S., Ha, Q.T., Collier, N.: An experiment in integrating sentiment features for tech stock prediction in twitter (2012)
32. Weng, B., Ahmed, M.A., Megahed, F.M.: Stock market one-day ahead movement prediction using disparate data sources. Expert Systems with Applications **79**, 153–163 (2017)