International Journal of Finance and Managerial Accounting, Vol.8, No.29, Spring 2023

Stock Price Forecasting with Support Vector Regression Based on Social Network Sentiment Analysis and Technical Analysis

Kamel Ebrahimian
Phd-studentat Department of Management، Faculty of Management and Accounting، Qazvin Branch، Islamic Azad University، Qazvin، Iran
Kamel61tb@gmail.com

Ebrahim Abbasi
Department of management, faculty of social sciences and economics, AlZahra University, Tehran, Iran.
corresponding author
abbasiebrahim2000@Alzahra.ac.ir

Akbar Alam Tabriz
Professor at Department of Industrial Management، Management and Accounting Faculty، Shahid Beheshti University، Tehran، Iran
Tabriz-a@sbu.ac.ir

Amir Mohammadzadeh
Department of Industrial at Department of Management، Islamic Azad University Qazvin، Qazvin، Iran
amn_1378@yahoo.com

Submit: 18/07/2021 **Accept:** 25/07/2021

ABSTRACT

For many years the following question has been a source of continuing controversy in both academic and business circles: To what extent can the past history of a common stock's price be used to make meaningful predictions concerning the future price of the stock (Fama, 1965). The rise of social networks and their role in daily life is significant due to the presence of these networks. Investors and their views on the stock market have affected financial markets. The purpose of this study is to predict the daily stock price using sentiment analysis and technical indicators. Therefore, support vector regression (SVR) is used in this study.

The tests involving feature combination with numeric and textual data and the proposed technical indicator features with the sentiment score series from tweets yield the best results of all, with classification accuracy for next day stock price prediction using the support vector regression model. The innovation of this study in comparison with other researches is stock price forecasting in short-term by combining the analysis of users' opinions and technical indicators, which uses the support vector regression.

The next section gives an overview of related work in the fields of text mining and stock price trend forecasting from unstructured data, methods and algorithms used in the research. In Section 3, the research methodology is stated and in the 4 and 5 sections, the results of modeling and discussion and conclusion will be stated.

Keywords: sentiment analysis, Vector Regression .

1. Introduction

For many years the following question has been a source of continuing controversy in both academic and business circles: To what extent can the past history of a common stock's price be used to make meaningful predictions concerning the future price of the stock (Fama, 1965). The rise of social networks and their role in daily life is significant due to the presence of these networks. Investors and their views on the stock market have affected financial markets. The purpose of this study is to predict the daily stock price using sentiment analysis and technical indicators. Therefore, support vector regression (SVR) is used in this study.
The tests involving feature combination with numeric and textual data and the proposed technical indicator features with the sentiment score series from tweets yield the best results of all, with classification accuracy for next day stock price prediction using the support vector regression model.

The innovation of this study in comparison with other researches is stock price forecasting in short-term by combining the analysis of users' opinions and technical indicators, which uses the support vector regression.

The next section gives an overview of related work in the fields of text mining and stock price trend forecasting from unstructured data, methods and algorithms used in the research. In Section 3, the research methodology is stated and in the 4 and 5 sections, the results of modeling and discussion and conclusion will be stated.

2. Literature Review

People are present in social networks to collaborate and participate with others for special or professional reasons. In fact, social networks are online communities that use the facilities to cultivate socialization and make users dependent on each other, and these networks are learning tools. Are in cyberspace.

Social networks provide an opportunity to connect with customers using richer and more accessible media. The interactive nature of these digital media not only helps sellers to share and exchange information with their customers, but also to share and also provides information exchange between customers (Malthouse, 2013).
Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.
Three steps involved are

- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.Text mining has become an exciting research field as it tries to discover valuable information from unstructured texts.The unstructured texts which contain vast amount of information cannot simply be used for further processing by computers.

Therefore, exact processing methods, algorithms and techniques are vital in order to extract this valuable information which is completed by using text mining.There are five basic text mining steps as under:

a) Collecting information from unstructured data.
b) Convert this information received into structured data
c) Identify the pattern from structured data
d) Analyze the pattern
e) Extract the valuable information and store in the database.

The third step is to form a matrix called the Term-Document-Matrix, in which the columns of the matrix are extracted words or terms and the rows of this matrix are comments or documents. In other words, this matrix indicates the frequency of each word in each document. (Text, comment, etc.). To get the results, a weighting algorithm called TF-IDF is used as in (1), which shows the importance of a term for a document. Suppose N is the total number of documents, n_i is the number of documents that Include the term i, D_{ji} is the frequency of the term i in the document j and D_j is the total number of terms in the document j.

$$T_iF - ID_jF = \frac{D_{ji}}{D_j} \log \frac{N}{n_i} \quad (1)$$

In step four, to determine the polarity of the comments in the two categories of positive and negative, label the

part of the comments that play the role of the target variable and then, by performing the first to third steps, the terms will be extracted. These terms will play the role of input variables.

Three hypotheses to advance the goal have been studied in this study

A: Hypothesis 1: There is a correlation between the volume of users' tweets and the volume of transactions the next day.

B: Hypothesis 2: The aggregation of daily emotions contains information for predicting stock price ratio (PRC).

C: Hypothesis 3: Is stock price forecasting using data mining of technical indicators significantly different from stock price forecasting using a combination of user emotion analysis and technical indicators?

Emotions could lead to short-term market fluctuations Schumacher and Chen (2009) assessed the impact of breaking news on stock prices twenty minutes after their release. Bollen et al. Used Twitter data to predict stock price trends (Bollen et al., 2011). They analyzed the emotions of ten million tweets using the GPOMS algorithm, and thus predicted the closing price of the Dow Jones Industrial Average with 87.6 %accuracy. Mittal and Goel based their work on a study by Bollen et al (2012). But with a larger set of views of ٤٠٠ million and the accuracy they obtained, it dropped from 87.6 %to 75% (Lee et al., 2014). In a study, Mashari et al. (2019) examined the predictability of short-term (bottom) and end (ceiling) short-term stock price trends using the Naïve Bayes model. Vatanparast et al. (2019) From a LM-BP neural network based on price time series introduced a method to predict stock price. Huang and Tsai presented a hybrid method utilizing filter-based feature selection, SOFM, and SVR, as a way to predict the stock price. Firstly, the filter-based feature selection was utilized to figure out crucial input technical indicators. After that, the SOFM was employed to cluster the training data into a number of disjointed clusters in ways that the components in every cluster are analogous. Lastly, the SVR was utilized to build an individual predicting approach for every cluster (Huang & Tsai, 2009). Their approach was showed via a case study on predicting the price of next day in the Taiwan stock exchange and the outcomes demonstrated that their strategy can enhance predicting precision and decrease the training time over the traditional single SVR.

3. Methodology

The present study is applied and this study examines the impact of social network users' feelings and technical indicators on stock price forecasting using the support vector regression algorithm.

3.1. Support vector regression

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs, is to recognize the function f (x) for training patterns x as in (2), which has the maximum margin of training values, in other words SVR It is a model that fits a curve with thickness ε to the data so that the least error is made in the test data (Vapnik, 1999).

Structural risk minimization is superior to the conventional method of trainrisk minimization used in algorithms such as neural networks and classical statistical methods, and does not converge to topical solutions for SVR. It is a function that is mapped to a real number based on training data from an input object. In regression problems, input vectors are mapped to a multidimensional space; a superplane is then created that separates the input vectors as far apart as possible. A kernel function is used to solve the problem of performing operations in large dimensions. In this case, the operation can be performed with the same speed as the input data space. In fact, using the kernel function, the problem of multidimensional and nonlinear mapping is solved. In fact, the purpose of the SVR is to estimate the parameters of weights and oblique functions that best fit the data, assuming that there is training data, if each input X has a number D attribute (in other words, belongs to a space with dimension D) and each point has a value of Y such that, like all regression methods, the goal is to find a function that relates to the input and output.

$$f(x,w) = w^T x + b \qquad (2)$$

To obtain the function f, it is necessary to calculate the values of w and b. To calculate the values of w and b as in (3) must be minimized (Vapnik, 1995).

$$R(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{l}\sum_{i=1}^{l} L_c(y_i, f_i(x,w)) \qquad (3)$$

Where C is a constant parameter and its value must be specified by the user. In fact, the function of the constant C parameter is to create balance and change the weights of the amount of the fine due to negligence (ε) and at the same time to maximize the size of the separation margin. The L_c function is a Vapnik function defined as in (4).

(4)

$$|y-f(x,w)|_{\varepsilon}\begin{cases} 0 & , \quad |y-f(x,w)| \leq \varepsilon \\ |y-f(x,w)| & -\varepsilon, \quad Otherwise \end{cases}$$

Equation (4) is rewritten as the maximization of (5)

(5)

$$\text{Max } L_p(a_i, a_i^*) = -\frac{1}{2}\sum_{j=1}^{l}(a_i - a_i^*)\left(a_j - a_j^*\right)\mathbf{x_i^T x_j} - \boldsymbol{\varepsilon}\sum_{j=1}^{l}(a_i - a_i^*) + \sum_{j=1}^{l}(a_i - a_i^*)y_i$$

Subject to constraints of (5) are as in (6).

$$\begin{cases} \sum_{j=1}^{l}(a_i - a_i^*) = 0 \\ 0 \leq a_i \leq C \ , i = 1, \dots, l \\ 0 \leq a_i^* \leq C \ , i = 1, \dots, l \end{cases} \quad (6)$$

By solving Eq.5, the SVR function, ie f, can be calculated as in (7) that using the kernel function:

(7)

$$f(x,w) = w_0^T x + b = \sum_{i=1}^{l}(a_i - a_i^*)\, x_i^T x + b$$

3.2. Model evaluation

Performance evaluation of the proposed approach is calculated based on various statistical parameters including the coefficient of determination of observational and estimated data (R^2) and Root Mean Square Error (RMSE) for training and traindata.

The R^2 and RMSE parameters for a variable such as X are defined according to (8) and (9).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)}{\sum_{i=1}^{n}(\bar{X} - \hat{X}_i)^2} \quad (8)$$

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2} \quad (9)$$

In these relations, X is an observed or measured variable, $\hat{X}$ is its estimated value of the variable.

3.3. Granger causality

One approach in examining the relationship between interacting variables is to look at the causality among these variables. Granger designed a statistical test, called the "Granger causality test," using a series of *t* tests and *F* tests to determine whether one time series is useful in predicting another time series (Granger, 1969). The Granger causality test does not necessarily address the cause-and-effect relation between variables as it may not indicate true causality. We assume that x_t and y_t are two stationary series; to determine whether x_t Granger causes y_t, first y_t is autoregressed on itself and the proper lag length is determined. In the next step the augmented autoregression of (11)

$$Y_t = \sum_{i=1}^{k}\alpha_i Y_{t-1} + \sum_{i=1}^{k}\beta_i X_{t-1} + \mu_i \quad (11)$$

Is estimated. The null hypothesis of "no Granger causality" is tested with a version of the *F* test, which checks whether all coefficients of x_{t-j}, namely β_{t-j}, are equal to 0. If all β_{t-j} are found to be equal to 0, then x_{t-j} does not precede y_{t-j}.

Geweke states that the validity of the test depends on the significance of both time series. That is, both time series must be meaningful. (Geweke, 1984)

3.4. Collecting, preparing and processing data

The statistical population is the data collected from 14 companies listed on the Tehran Stock Exchange. The textual data of the research are the daily comments and tweets of the users of Sahamyab website. Complete list of the initially collected properties can be seen in Table (1).

Modeling is based on 20 technical indicators of daily prices (Table 2) and one emotion analysis variable. Data from April 2016 to the end of March 2017 are used for modeling.

Table (1): Tweet Features

Features	Description
Name, username, comment, date, time of publication, source of message	Comments
Name, username, comment	Users

Table (2): Technical indicators

Source	Variable description	Indicator
Pring,1991	Know Sure Thing	KST
Kaufman,2013	Kaufman's Adaptive Moving Average	KAMA
Coppock,1969	Coppock Curve	Coppock Curve
Williams,2014	Williams %R	R%
Botes & Douglas,2010	Vortex Indicator	VI
Morphy,*2010*	Decision Point Price Momentum	DPPM
Williams ,1985	Ultimate Oscillator	UO
Gurib ikhlas ,2018	Average Directional Index	ADI
Sakarya & Deng ,2013	Money Flow Index	MFI
Lambert 1980	Commodity Channel Index	CCI
Balu,1991	True strength index	TSI
Murphy,2009	relative strength index	RSI
Dizart,1983	Negative Volume Index	NVI
Niman,2009	Chalkin Money Flow	CMF
Derrosi 1983	MASS Index	MASS
Hussein,2018	Price Rate of Change	PRC
Wilder,1987	Average True Range	ATR
Antakaris,2019	percentage price oscillator	PPO
Chande,1994	STOCHASTIC RSI	StochRSI
Person2004	Stochastic oscillator	D%

3.5. Research method

3.5.1. Data processing

The companies along with the number of comments collected April 2016 to the end of March 2017 from the website of the stock finder is given in Table (2). The total number of tweets in these 14 Tickers was 250.674 comments. And reduced to 234.403 comments by removing duplicate comments.To pre-process user comments, the following steps are taken:

A: Duplicate comments and comments on weekends and public holidays and stop time of the Ticker were removed.

B: In the comments, the @ sign was replaced with the text "Etsain."

C: Ticker of each company was replaced with the word "share Ticker" in the comments.

TSE Client, AmoBroker and Excel software were used to obtain the monitors through programming, and R programming language was used for data mining and text mining.

3.5.2. Tagging user comments

In this study, instead of using emotional dictionaries, polarity analysis of investors' feelings is used because although in determining the feelings of tweets, subjectivity and objectivity are considered, but in many cases, neutral (objective) tweets in the situation Normal should be considered as polar tweets in this research. For example, the comment "#car is no news, just a waste of time" will most likely be marked as neutral and neutral sentences Polarity analysis is out, when in fact it shows that a user expects the stock price to fall. The problem is that we used the tagging of some of the tweets with the help of experts to teach the classification algorithms

Table (3): some user comments and their tags

Label	Argument	Comment
Positive	A short look with positive feelings towards the #KHODRO Ticker that will go through a growing trend	#KHODRO buy tomorrow until 11 o'clock, whatever institutions it offers, which you can not buy later.
Neutral	Without expressing any emotion, the user asks about the impact of the projects on the current situation and how it differs from the past.	#Kayson has not had a small project in the past, so why hasn't it had any effect on its stock price before?
Negative	The user expresses his / her opinion about the false growth of the share, and suggests the replacement of the share.	#Shepley, this 100 Toman share went up with speculation Change your stock… Do not get married

3.5.3. Modeling

In the first part of the research, a model is developed to classify opinions into two categories, positive and negative, and after teaching the model and determining the complete class of comments, the daily emotion aggregation variable is calculated. This variable is derived from the algebraic sum of people's daily emotions and is the difference between the number of negative comments and the number of positive comments of the relevant stocks. In other words, if i is the name of one of the Company in Table (۴), then the aggregation of the feelings of the j day of these Tickers is obtained as in (12)

$$S_{ij} = S_{ij}^{+} - S_{ij}^{-} \quad (12)$$

where S_{ij}^{+} number of positive comments on the stock i on day j, S_{ij}^{-} the number of negative comments on the same stock on day j and S_{ij} is the aggregation of emotions and the probability is going to increase the next day to move the stock and if $S_{ij}<$ ۰ then the opposite of the above can be established.

The second part of the analysis is related to correlation, causality, and the third part is related to modeling with a combined approach based on technical indicators and emotion analysis to predict stock prices. Therefore, in this section, there are 21 daily input variables for a Ticker. 20 Variable Related to daily technical indicators and another variable is the aggregation of daily emotions of those stocks.Variable The goal in this section is stock price.

Statistics of the number of comments published by users, the number of trading days, the rank of followers on the stock exchange site and its average daily visits are described in Table(4).

Table (4): List of stocks with number of comments

Average Daily Visits	Rank Follower	Number of Trading Day	Comments	Company Name	Average Daily Visits	Rank Follower	Number of Trading Day	Comments	Company Name
45581	6	435	20844	KHODRO	2450	65	435	4726	AKHABER
2730	117	472	21123	KEYSON	3851	49	349	10252	HAFFARI
20624	5	461	22379	KHESAPA	4523	29	436	11424	FARAK
21855	2	356	26912	VATEJARAT	18166	13	445	12139	SHETRAN
7275	4	463	24219	ZOB	11295	12	461	14337	VABESADER
3239	18	458	22012	TEPCO	10461	10	460	14368	SHEBANDAR
4778	23	436	28015	SHEPLAY	2416	62	453	17924	SETRAN

4. Results

Three types of analysis were examined in this study: the first analysis is related to the natural language processing and emotion analysis approach; the second analysis is related to correlation and causality of Granger and the third analysis is related to the combined approach of stock price forecasting.

4.1. Natural language processing and emotion analysis

A random sample of 2,344 tweets was selected as train and test dataset, which were manually determined by financial market experts, their polarity in both positive and negative categories. From the total sample, 1644 tweets were selected as a training and 704 tweets as a test sample. The results of the classification algorithms describe the use of unigrams (single words) and bigrams (two words) to construct attributes (variables). Simple Bayesian algorithms, decision trees and support vectors to construct the class prediction model the comments used and their evaluation results for the training and test dataset are given in Table(5) and table(6), respectively. We reached 80/37% accuracy for calling test tweets.

There are similar results in evaluating trainsamples with this algorithm in Table (6). According to this table, the support vector machine algorithm has an accuracy of 70.20%. Other comments will be used.

After tagging, the views of the daily volume volume variables, the number of daily positive comments, the number of daily negative comments, and the aggregation of daily emotions per share were calculated.

Table (5): Evaluation results of Test sample classification algorithms (emotion analysis)

Cunfosion Matrix			Recall Negative tweets	Recall positive tweets	accuracy	algorithm
	pos	neg	83/44	77/33	80.37	SVM
pos	638	187				
neg	135	680				

Table (6): Evaluation Results of train Sample Classification Algorithms (Emotion Analysis)

Cunfosion Matrix			Recall Negative tweets	Recall positive tweets	accuracy	algorithm
	pos	neg	67/47	72/53	70/20	SVM
pos	272	103				
neg	107	222				

4.2. Test the first hypothesis

The important fundamental question is whether the volume of x Ticker tweets and the volume of transactions of that Ticker the next day are related? And can we expect that there is a significant correlation between the volume of tweets of each Ticker and the volume of transactions the next day? The value of correlation coefficient and p-value related to the correlation hypothesis test in Table (7) indicates that in all Tickers except Khesapa and Kayson, there is a positive and significant correlation between the total volume of tweets and the volume of next day transactions. The volume of positive tweets and the volume of transactions the next day are significant only in 9 Tickers, and finally, there is no significant correlation between the volume of negative tweets and the volume of transactions the next day.

4.3. Test the second hypothesis

Given the correlation between the volume of tweets and the volume of transactions the next day, can we go one step further and ask the question, or does the aggregation of daily emotions have information to predict the stock price change ratio (PRC)? To answer this question, the Granger causality test should be used, but before that, the Dickey-Fuller test is necessary to examine the significance of both time series. It is important to note that both series must be delayed by a maximum of one. For further explanation and assumptions, suppose Granger causality of today's emotion aggregation (S_i) to predict tomorrow's price change PRC_{i+1} is defined by the vector autoregressive model as in (13).

$$PRC_{i+1} = \alpha_0 + \alpha_1 PRC_i + \beta_0 + \beta_1 S_i \quad (13)$$

Therefore, both time series must have a maximum delay of one mean, and if Tickers can not be meaningful in either Dickey-Fuller or Granger causality tests, they will be removed from the analysis. Fuller is significant, so both the PRC and S variables have a mean lag of one.

According to Table (9), the Granger cause test is meaningful in all Tickers except Tepco and Vatejarat, so these two Tickers will be removed from further analysis

Table (8): Correlation between the volume of tweets and the volume of transactions the next day

Ticker	Negative tweets		Positive Tweets		Tweets	
	P-value	R	P-value	R	P-value	R
AKHABER	0/2528	۰/۰۵۵۰	۰/۰۰۰۰	۰/۱۹۵۰	۰/۰۰۰۰	۰/۲۵۹۰
TEPKO	0/1708	۰/۰۷۳۶	۰/۰۲۳۳	۰/۱۲۲۰	۰/۰۰۰۱	۰/۲۱۴۰
VATEJARAT	0/4396	۰/۰۳۷۲	۰/۰۰۳۵	۰/۱۴۰۰	۰/۰۰۰۰	۰/۱۹۴۰
HAFFARI	0/8213	۰/۰۱۰۷ -	۰/۰۰۰۰	۰/۱۹۳۰	۰/۰۰۰۰	۰/۲۰۰۰
KHESAPA	0/1781	۰/۰۶۲۹	۰/۵۶۰۲	۰/۰۲۷۲	۰/۰۵۴۱	۰/۰۸۹۹
KHODRO	0/0860	۰/۰۸۱۶	۰/۰۷۸۷	۰/۰۸۲۲	۰/۰۰۰۲	۰/۱۷۳۰
ZOB	0/1050	۰/۰۷۶۴	۰/۴۹۱۸	۰/۰۳۲۴	۰/۰۳۰۴	۰/۱۰۲۰
SETRAN	0/1011	۰/۰۷۸۸	۰/۱۰۴۹	۰/۰۷۸۰	۰/۰۰۰۶	۰/۱۶۵۰
SHETRAN	0/9610	۰/۰۰۲۲۶	۰/۰۰۰۱	۰/۱۷۶۰	۰/۰۰۰۰	۰/۲۰۹۰
SHEBANDAR	0/2076	۰/۰۵۸۹	۰/۰۰۶۳	۰/۱۲۷۰	۰/۰۰۰۰	۰/۱۹۶۰
SHEPLAY	0/8316	۰/۰۱۱۳	۰/۰۰۴۲	۰/۱۵۲۰	۰/۰۰۰۳	۰/۱۹۲۰
FARAK	0/2211	۰/۰۵۷۰-	۰/۰۰۰۰	۰/۱۹۹۰	۰/۰۰۰۱	۰/۱۷۷۰
KEYSON	0/7866	۰/۰۱۲۷	۰/۱۹۰۹	۰/۰۶۱۳	۰/۱۱۸۹	۰/۰۷۳۰
VABESADER	0/5312	۰/۰۳۰۱	۰/۰۰۴۸	۰/۱۳۵۰	۰/۰۰۰۳	۰/۱۷۳۰

Table (8): Dickey Fuller test results

Aggregation of emotions		PRC		Ticker	Aggregation of emotions		PRC		Ticker
P-value	Dickey Fuller	P-value	Dickey Fuller		P-value	Dickey Fuller فولر	P-value	Dickey Fuller	
۰/۰۱*	-۱۴/۴۶	۰/۰۱*	-۱۰/۵۵	SETRAN	۰/۰۱*	-۱۴/۷۲	۰/۰۱*	-۱۲/۸۸	AKHABER
۰/۰۱*	-۱۷/۱۸	۰/۰۱*	-۱۴/۱۲	SHETRAN	۰/۰۱*	-۱۱/۲۲	۰/۰۱*	-۱۲/۹۳	TEPKO
۰/۰۱*	-۱۵/۵۹	۰/۰۱*	-۱۵/۲۵	SHEBANDAR	۰/۰۱*	-۱۵/۱۹	۰/۰۱*	-۱۴/۵۱	VATEJARAT
۰/۰۱*	-۱۳/۰۸	۰/۰۱*	-۱۲/۹۰	SHEPLAY	۰/۰۱*	-۱۵/۴۹	۰/۰۱*	-۱۲/۵۰	HAFFARI
۰/۰۱*	-۱۷/۴۱	۰/۰۱*	-۱۳/۹۴	FARAK	۰/۰۱*	-۱۴/۵۲	۰/۰۱*	-۱۲/۴۹	KHESAPA
۰/۰۱*	-۱۵/۰۱	۰/۰۱*	-۱۴/۵۱	KEYSON	۰/۰۱*	-۱۵/۱۰	۰/۰۱*	-۱۳/۰۶	KHODRO
۰/۰۱*	-۱۵/۰۷	۰/۰۱*	-۱۴/۸۰	VABESADER	۰/۰۱*	-۱۵/۵۴	۰/۰۱*	-۱۴/۳۴	ZOB

* Less than 0.01

Table (9): Test results for Granger causality, emotion aggregation to predict price change ratio

TiCK	F	P-value	TiCK	F	P-value
AKHABER	2.59	0.000	SHEPLAY	33.7	0.000
TEPKO	0.84	0.358	KHESAPA	19.02	0.000
VATEJARAT	0.39	0.529	KHODRO	50.12	0.000
HAFFARI	2.54	0.000	ZOB	21.39	0.000
SETRAN	34.3	0.000	FARAK	64.11	0.0000
SHETRAN	41.6	0.000	KEYSON	26.65	0.0000
SHEBANDAR	8.11	0.0046	VABESADER	7.45	0.0063

4.4. Stock price forecast

To predict stock prices using support vector regression in two modes, using Technical Analysis (TA) and combining Technical Analysis and Sentiment Analysis (TASA) and using RMSE criteria and coefficient of determination, the two models will be compared. In fact, the coefficient of determination shows how much of the change in the response variable is explained by the independent variables. It should be noted that here the response variable is the daily closed stock price.For predictive models, 70% of the total samples will be used for model training and the remaining 30% will be used for model testing. In regression methods,

independent variables should not be in line with each other. This means that independent variables should not be a function of each other and should be statistically independent of each other. Otherwise, the regression model has a challenge called linearity and its results are unreliable. The measure used for alignment is the variance inflation index, which is represented by VIF. James et al. (2014) stated that if the value of this index for a variable is more than 5 or 10, that variable has a line with other variables. There are two ways to fix the alignment, either the variable can be removed from the model or exploratory factor analysis can be used. In the analysis of the main factors, an attempt is made to make a factor from the variables that have a high correlation with each other. It is noteworthy that the constructed factors are not in line with each other and are completely independent of each other. Table (10) shows the VIF values of the research variables. According to this table, it is clear that the alignment between the researches variables is high, so exploratory factor analysis will be used to continue the work.

Table (11) shows that the technical variables are divided into ۴main factors and the factor load of each variable in each factor is specified. Used to use the SVR algorithm.

4.5. Test the third hypothesis

The third hypothesis is whether stock price forecasting using data mining of technical indicators is significantly different from stock price forecasting using a combination of user emotion analysis and technical indicators?

RMSE values and coefficients of determination of both models (TA) and (TASA) are reported in Table (۱۲) for training samples. According to this table, the model (TA) has a lower RMSE and a higher coefficient of determination, and this shows that the variable of aggregation of daily emotions has no effect on stock price forecasting. The difference is that the aggregation of daily emotions has a positive effect on the direction of stock movement, but the same variable can not be effective in predicting stock prices, and Table (۱۳) shows similar results. The issue here is not the analysis of emotions but the prediction of stock prices using a very simple technique whose results clearly show the power of this prediction. The diagrams in Fig.1 actually show the predicted values versus the actual values of the stock package price, which is similar to a straight line in each model, indicating that the actual values are very similar to the original values.

Table (10): VIF index values of technical variables

(Constant)	Tolerance	VIF	(Constant)	Tolerance	VIF
PPO	0.063	15.94	PRC	0.532	1.88
RSI	0.021	47.624	ADX	0.335	2.989
STOCHASTIC	0.076	13.16	CCI	0.122	8.215
STOCHASTIC16	0.191	5.244	CPPC	0.155	6.451
TSI	0.026	37.971	ATR	0.073	13.609
UO	0.164	6.111	KAMMA	0.076	13.104
VI	0.191	5.24	KST	0.047	21.102
CMF	0.313	3.195	MASS	0.383	2.613
PMO	0.033	30.569	MFI	0.285	3.508
MACD	0.394	2.536	NVI	0.767	1.304

Table(11): Factors generated by technical variables based on Exploratory factor analysis

Indicators	FACTORS			
	1	2	3	4
WILLIAMSR	۰/۹۳			
STOCHASTIC	۰/۹۳			
STOCHASTIC16	۰/۸۸			
CCI	۰/۸۵			
UO	۰/۷۴			
RSI	۰/۶۹			
PRC	۰/۶۴			
VI	۰/۶۱			
PMO		۰/۹۱		
PPO		۰/۹		

Indicators	FACTORS			
	1	2	3	4
KST		۰/۹		
TSI		۰/۸۲		
MASS		۰/۷۸		
CPPC		۰/۶۸		
ADX			۰/۷۶	
MFI			۰/۶۲	
CMF			۰/۵۷	
KAMMA				۰/۹۳۹
ATR				۰/۹۱۸
NVI				۰/۴۸۵
MACD				۰/۴۸
VOUME				۰/۴۸-

Table (۱۲): Root Mean Square Error (RMSE) and coefficient of determination (R2) for training samples

Model	RMSE	R2
TA	۱۰۱۰/۶۹۴	۰/۹۷۲۵
TASE	۱۱۲۶/۴۸۲	۰/۹۶۶۴

Table (۱۳): Root Mean Square Error (RMSE) and coefficient of determination (R2) for test sample

Model	RMSE	R2
TA	۱۱۳۵/۷۷۹	۰/۹۵۲۵
TASE	۱۳۹۶/۱۰۱	۰/۹۲۷۸

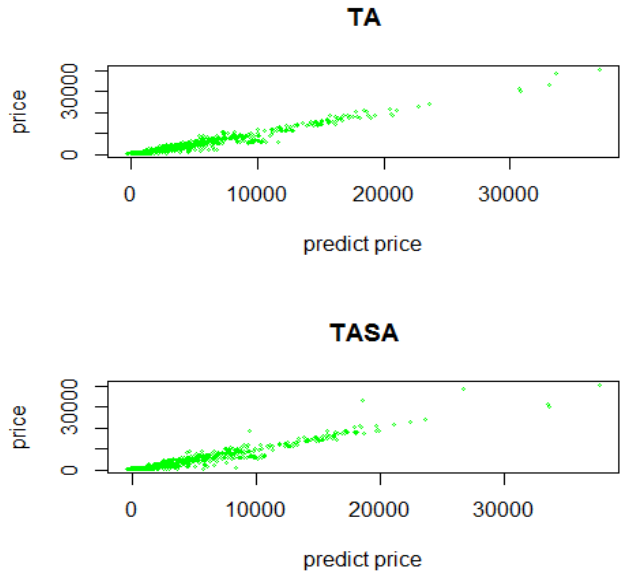

Fig (۱): Predicted price chart versus actual stock price using Technical Analysis (TA) and Combining Technical Analysis with Sentiment Analysis (TASA)

5. Discussion and Conclusion

In this study, using a combined approach of emotion analysis and technical indicators, a model for stock price forecasting using data mining and support vector regression was developed. According to the research findings, stock price forecasting using support vector regression with input the technical indicator data, considering the RMSE criterion as well as the coefficient of determination, has a higher accuracy than the time when it simultaneously has the data of technical indicators and the aggregation of users' emotions in social networks. From the country, including researches (Lee et al., 2014) and (Koi, Roy, Winston, 2013) and domestic research (Rai et al., 2016), it is compatible that the ability of the social network to predict stock prices compared to predict changes. The stock price (direction) is higher.

The major systems for market prediction based on online text mining have been reviewed and some of the predominant gaps that exist within them have been identified. The review was conducted on three major aspects, namely: pre-processing, machine learning and the evaluation mechanism; with each breaking down into multiple sub-discussions. It is believed to be the first effort to provide a comprehensive review from a holistic and interdisciplinary point of view. This work intended to accomplish: Firstly, facilitation of integration of research activities from different fields on the topic of market prediction based on online text mining; Secondly, provision of a study-framework to isolate the problem or different aspects of it in order to clarify the path for further improvement; Thirdly, submission of directional and theoretical suggestions for future research.

In this study, a wide range of opinions and feelings of online traders and its effect on stock prices and mass behavior of investors were examined, while in previous studies, only news or rumors were used to predict, as well as the behavior of small traders in forecasting. Price is important, but the asymmetric effects of major shareholders can be effective as a result of the model.

The objectives of the study have been achieved by having:

- Demonstrated that Support Vector Machine is the best classifier with the overall accuracy rates of 70.20%.
- Discovered that users' activity sahamyab significantly positively correlates to the stock trading volume the next business day.
- Determined that simply summed up collective sentiments has Powerful prediction on the change of stock price for the next day in 12/14 of the stocks studied by using Granger Causality test.
- discovered that the overall accuracy rate of predicting price of stocks by using the technical analysis is better than combining Sentiment analysis and technical indicators

As a future research, some variables such as Fundamental factors that affect the value of companies, macroeconomic variables such as interest rates, budgets, and global price changes and their effect on user feedback can also be added to the model, using more data and more user feedback. The characteristics and level of ability of users and the use of graph analysis and online networks between users and their level of knowledge in investment, as well as the development of a glossary for the capital market in the field of emotion analysis, should be researched and future research. It also used artificial neural networks to optimize the text mining process and increase prediction accuracy.

References

1) Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, Vol. 2, no. 1, pp.1-8.
2) Fama, E. F. (1965). The behavior of stock-market prices. The journal of Business, Vol.38, no. 1, pp. 34-105.
3) Geweke, J. (1984). Inference and causality in economic time series models. Handbook of econometrics, Vol.2, pp.1101-1144.
4) Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, pp.424-438.
5) Huang, C. L., & Tsai, C. Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Systems with applications, Vol.36, no. 2, pp. 1529-1539.
6) Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014, May). On the Importance of

Text Analysis for Stock Price Prediction. In LREC (Vol. 2014, pp. 1170-1175).

7) Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: Introducing the social CRM house. Journal of interactive marketing, Vol: 27.No. 4, pp.270-280.
8) Mittal، Anshul، and Arpit Goel. (۲۰۱۲). Stock Prediction Using Twitter Sentiment Analysis. Stanford University، CS۲۲۹. Available online: http://cs۲۲۹.stanford.edu/proj۲۰۱۱/GoelMittalStockMarketPredictionUsingTwitterSenti
9) Moshari, M., Didehkhani, H., Khalili Dameghani, K., & Abbasi, E. (2019). Designing a Hybrid Intelligent Model for Prediction of Stock Price Golden Points. Journal of Investment Knowledge, Vol. 8, no. 29, pp.45-66.
10) Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), Vol.27, No. 2, pp.1-19.
11) Vapnik, V. N. (1995). The nature of statistical learning theory (No. 04; Q325. 7, V3.).
12) Vapnik, V. (1999). The Nature of Statistical Learning Theory. Springer Science & Business Media.
13) Vatanparast, M., Asadi, M., Mohammadi, S., & Babaei, A. (2019). Stock price prediction based on LM-BP neural network and over-point estimation by counting time intervals: Evidence from the Stock Exchange.، Vol:۱۰، No. ۳۹; pp. ۱۹۳- ۲۱۸.