IGI PUBLISHING
WWW.IGI-GLOBAL.COM

International Journal of Business Intelligence Research

October-December 2013, Vol. 4, No. 4

Table of Contents

Research Articles

1 **Correlation between the Economy News and Stock Market in Turkey**
Sadi Evren Seker, Department of Computer Engineering, Istanbul University, Istanbul, Republic of Turkey
Cihan Mert, Department of Electrical Engineering, University of Texas at Dallas, TX, USA
Khaled Al-Naami, Department of Computer Science, University of Texas at Dallas, TX, USA
Nuri Ozalp, Turkish Science Foundation, Istanbul, Republic of Turkey
Ugur Ayan, Turkish Science Foundation, Istanbul, Republic of Turkey

22 **Propose a Conceptual Model of Adaptive Competitive Intelligence (ACI)**
Sareh Mohammadalian, Faculty of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, Iran
Eslam Nazemi, Faculty of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, Iran
Mohammad Jafar Tarokh, Faculty of Industrial Engineering, Khajeh Nasir Toosi University, Tehran, Iran

33 **A Hierarchy of Metadata Elements for Business Intelligence Information Resource Retrieval**
Neil Foshay, Department of Information Systems, St. Francis Xavier University, Antigonish, Canada
Todd Boyle, Department of Information Systems, St. Francis Xavier University, Antigonish, Canada

45 **Determinants of Process Change Outcome: An Exploratory Case Study Research Model**
Chelsey Hill-Esler, Department of Decision Science, Drexel University, Philadelphia, PA, USA

61 **Towards Automation of Business Intelligence Services Using Hybrid Intelligent System Approach**
R. M. Sonar, Shailesh J Mehta School of Management, Indian Institute of Technology, Mumbai, India

The *International Journal of Business Intelligence Research* is indexed or listed in the following: Bacon's Media Directory; Cabell's Directories; DBLP; Google Scholar; INSPEC; Library & Information Science Abstracts (LISA); MediaFinder; The Standard Periodical Directory; Ulrich's Periodicals Directory

Correlation between the Economy News and Stock Market in Turkey

Sadi Evren Seker, Department of Computer Engineering, Istanbul University, Istanbul, Republic of Turkey

Cihan Mert, Department of Electrical Engineering, University of Texas at Dallas, TX, USA

Khaled Al-Naami, Department of Computer Science, University of Texas at Dallas, TX, USA

Nuri Ozalp, Turkish Science Foundation, Istanbul, Republic of Turkey

Ugur Ayan, Turkish Science Foundation, Istanbul, Republic of Turkey

ABSTRACT

Depending on the market strength and structure, it is a known fact that there is a correlation between the stock market values and the content in newspapers. The correlation increases in weak and speculative markets, while they never get reduced to zero in the strongest markets. This research focuses on the correlation between the economic news published in a highly circulating newspaper in Turkey and the stock market closing values in Turkey. In the research several feature extraction methodologies are implemented on both of the data sources, which are the stock market values and economic news. Since the economic news is in natural language format, the text mining technique, term frequency – inverse document frequency is implemented. On the other hand, the time series analysis methods like random walk, Bollinger band, moving average or difference are applied over the stock market values. After the feature extraction step, the classification methods are built on the well-known classifiers support vector machine, k-nearest neighborhood and decision tree. Moreover, an ensemble classifier based on majority voting is implemented on top of these classifiers. The success rates show that the results are satisfactory to claim the methods implemented in this study can be spread to future research with similar data sets from other countries.

Keywords: *Classification, k-Nearest Neighbors Algorithm (kNN), Sentimental Analysis, Stock Market Analysis, Support Vector Machines (SVM), Text Mining, Time Series Analysis*

1. INTRODUCTION

This study is built on two datasets. The first one is the economic news from one of the high-circulating newspapers in Turkey, which have special pages for economic news. We have collected only economic news, which are separate from other news like sports or local news etc. The second dataset is the stock market closing values and we have applied several time series

DOI: 10.4018/ijbir.2013100101

analyses over it. The properties of the dataset will be explained in the experiments section. We have processed the news text via the text mining approach called term frequency – inverse document frequency (TF-IDF) which will be explained in the methodology section. On the other hand, we have processed the stock market closing values by using time series analysis, which are random walk (RW), Bollinger band (BB), relative strength index (RSI), moving average convergence / divergence (MACD), momentum, rate of change (ROC), acceleration, difference and their variations. Finally we have investigated the correlation between these two feature vectors by using the support vector machines (SVM), k-nearest neighborhood (KNN) and decision tree-based classification (C4.5). Furthermore an ensemble classification based on majority vote learning (MaVL) is implemented on top of these three classifiers, which is discussed in the section of classification. Also in this paper, we discuss the implementation details and the methodology of evaluation over the classification results, in the evaluation section.

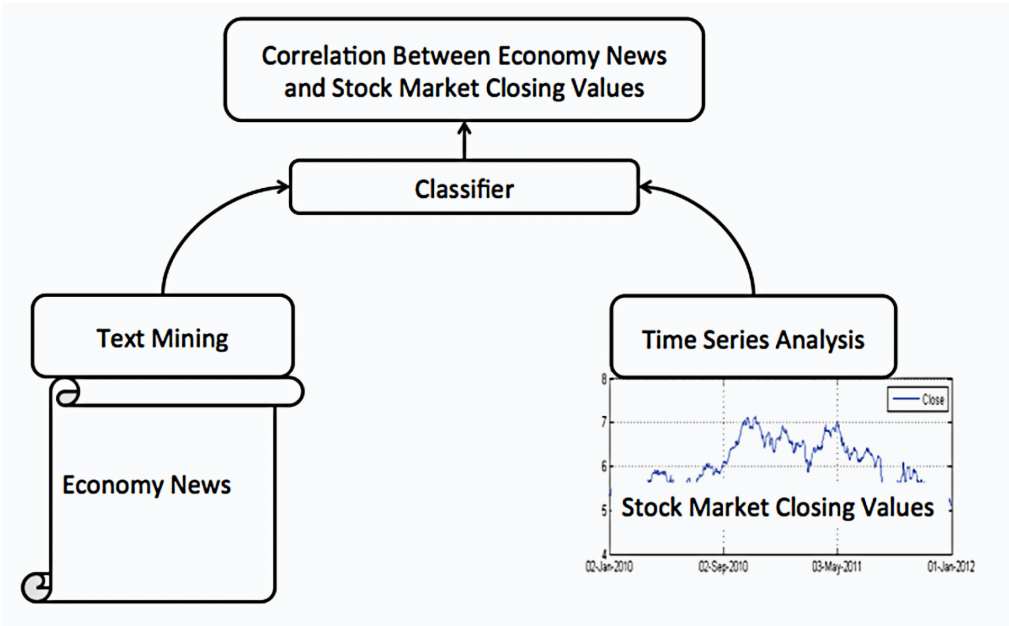
2. PROBLEM STATEMENT

We have two datasets, the economy news and stock market closing values, and we have applied a layered approach in this study, as illustrated by Figure 1. At the bottom level starting from the datasets, we build a feature extraction methodology. For the economic news, which is in natural language, we have applied sentimental analysis via the text mining methodologies. On the other hand, we have applied the random walk method for the feature extraction from the stock market closing values.

The correlation between news and the stock market is one of the indicators of the speculative markets (Nikfarjam, Emadzadeh, & Muthaiyah, 2010).

One of the difficulties in this study is dealing with the natural language data source, which requires a feature extraction. The other difficulty is dealing with a stock market value, which is considered as a signal. The size of data we are dealing with, which can be considered big data, is also problematic. The dataset holds 131,248 distinct words and when the feature vector of

Figure 1. Bottom up overview of study



each economic news item is collected, the total size of the feature vector is beyond 2.5 GByte, which is beyond the computation capacity of a single computer with these classification algorithms. For example, only one of the classification algorithms, we will call SVM requires slightly more memory than 1TByte in this case.

3. RELATED WORK

Research about stock markets has always been an interesting area of study because of its impact on the business and financial world. Furthermore, the developments in the computer science field have opened a door to research on the stock markets parallel to text-based news after 2000s. Most of the works use text mining tools over the news. For example, one of the popular methods is to use the bag of words (2-6) where some use local dictionaries (2,6) or some use some text mining tools like IBM Text Miner(Fung, Yu, & Lam, 2002). Others use the TF-IDF approach (3,6), some predefined term dictionaries(Rachlin, Last, Alberg, & Kandel, 2007) or part of speech tagging (Mahajan, Dey, & Haque, 2008) and concept maps (Soni, Eck, Jan, & Uzay, 2007). Also classification varies from latent Drichlet allocation (Mahajan, Dey, & Haque, 2008) to SVM (3-8) or decision tree (Rachlin, Last, Alberg, & Kandel, 2007). The success rates vary from 45% to 82% and the data sources are Reuters market 3000 extra (Fung, Yu, & Lam, 2002), PRNewswire (Mittermayer & Knolmayer, NewsCATS: A News Categorization and Trading System, 2006), FT Intelligence (Soni, Eck, Jan, & Uzay, 2007), Australian Financial Review(Halgamuge, Y, & Hsu, 2007), Forbes and Reuters web sites (Rachlin, Last, Alberg, & Kandel, 2007), Yahoo Finance (Schumaker & Chen, 2009) or Wall Street Journal (Mahajan, Dey, & Haque, 2008).

While there is a great amount of published articles about machine learning, time series, other classification and regression methods in prediction of stock market prices, the numbers of articles involving the application of finding correlation between stock market and daily news using text mining techniques are low. News, especially economic news based stock market prediction, can be considered as a text classification/mining task. The main aim is to predict some aspects of the stock market such as price or volatility based on the news content or derived text features. Based on a forecasting goal, a set of final classes are defined, such as "Up" (increase in price), "Down" (decrease in price), "Balance" (no change in price), etc. All proposed algorithms are supposed to classify the incoming news into these classes.

One of the first works on forecasting stock market uses hints (statistics, ratios and interpretations of trend-charting techniques) listed by domain experts to predict weekly UP's and DOWN's of the stock market (Braun, 1987). Generally, published related works are built around a central learning algorithm (mostly a classifier) for predicting the sentiment (price direction) of news articles. Wutrich et al. (1998) try to predict stock market indices using information contained in articles published in The Wall Street Journal (www.wsj.com), Financial Times (www.ft.com), Reuters (www.investools.com), Dow Jones (www.asianupdate.com), and Bloomberg (www.bloomberg.com) based on k-NN learning algorithm and the Euclidean similarity measure.

Similarly, Mittermayer (2004) tries to predict price trends (incline, decline or flat) immediately after press release publications. "Good news" articles are categorized as incline if the stock price relevant to the given article has increased with a peek of at least +3% from its original value at the publication time at some point during the 60 minutes that follows. The average price level in this period has to be at least 1% above the price at publication. For "Bad news" the opposite is true; the movements have to be in the negative direction instead. The rest of the articles in between are classified as "No movers".

Schumaker and Chen (2006) studied the effect of financial news articles on three different derived text mining features, Bag of Words, Noun Phrases, and Named Entities, and their

ability to guess discrete number stock prices after an article release, which is a regression problem and not a categorization problem. In this study, Named Entities features scheme gives better results than the Bag of Words on 1-class SVM algorithm.

Mittermayer and Knolmayer (2006) proposed an automated text categorization system using a hand-made thesaurus to forecast intraday stock price trends from information contained in press releases. This system includes three engines: the document preprocessing Engine automatically preprocesses incoming press releases, the categorization engine sorts the press releases into different categories and, finally, the trading engine triggers trading recommendations for the corresponding security. The system creates the feature list with the bag of words method using either trivial functions like Collection Term Frequency (CTF), Inverse Document Frequency (IDF), and CTFxIDF or ATC-specific functions like Chi-squared (CHI), Information Gain (IG), and Odd's Ratio (OR).

Falinouss (2007) developed a methodology which uses some time series segmentation techniques to find stock price trends by reducing the noise in the price curve. Two different methods of labeling the news articles were introduced. Method one is "observed time lag", which has a time lag between publishing time and when the public absorbs and acts on it. The second method is built on the efficient market theory, which states that there is no time lag between publishing time and market correction for this piece of information.

Schumaker and Chen (2010) proposed another learning system named the Arizona Financial Text System that uses financial news articles and stock quotes as its input to predict feature stock price movements. This system uses proper nouns that occur three or more times to be included in the feature set. It distinguishes itself by using a support vector regression (SVR) algorithm instead of a classification in order to predict a future price value when given an article.

Tan (2010) describes a classifier based on labeling by trading volume or by returns that can only predict whether a news article will initiate a high number of trades or not; it cannot predict the direction of the price movement. The division of the volume labels is controlled by whether the given volume is above or below the average trading volume. Labeling by returns looks for abnormal return amounts and labels news articles from that.

Our aim is to put a new dataset into the literature and to study the correlation between the new dataset and stock market values in Turkey where further research can be placed for the market structure and strength in the future.

4. BACKGROUND

We have implemented TF-IDF and RW methods as already explained in the introduction; this section will discuss these methods in detail. Also one of the difficulties is the number of words we are dealing with. We have implemented the information gain calculation for eliminating some of the features.

4.1. Term Frequency: Inverse Document Frequency

For the TF-IDF calculation is given in Equation 1:

$$tfidf\left(t,d,D\right)=tf\left(t,d\right)x\,idf\left(t,D\right) \tag{1}$$

where t is the selected term, d is the selected document and D is all documents in the corpus. Also TF-IDF calculation in above formula is built over term frequency (TF) and inverse document frequency (IDF), which can be rewritten as in Equation 2:

$$tf\left(t,d\right)=\frac{f\left(t,d\right)}{\max\left\{f\left(w,d\right):w\in d\right\}} \tag{2}$$

where f is the frequency function and w is the word with maximum occurrence. Also the formulation of IDF is given in Equation 3:

$$id(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

where |D| indicates the cardinality of D, which is the total number of documents in the corpus.

4.2. Financial Time Series Analysis

The stock market closing values are collected from the web page of the Istanbul Stock Market which is available for public download and usage.

As demonstrated in Figure 2, the values are in local currency, which is Turkish Lira that we will ignore. This section focuses on the time series analysis methods. We will try to briefly explain the methods and their variations.

4.2.1. Random Walk

We use a relative feature extraction method and get the difference between two consecutive closing values like for any $C_i \in C$ where C is the set of closing values, we collect pairs of $<C_t, C_{t-1}>$ for each news where t is the publication date of the news.

The collected features then subtract from each other to see if there is an increase or decrease in the closing value of the stock market as in Equation 4:

$$Feature = C_t - C_{t-1} \quad (4)$$

If the feature value of the closing values of the stock market is positive (+), we consider it as 1; if the value is negative (-),we consider it as -1. Also there are some dates which do not hold the closing values such as weekends. For those dates we consider the class label of the news as 0.

This approach can be considered as random walk in the literature. The formulation of random walk is given as in Equation 5 (Lu, Huang, Zhang, & Chen, 2009):

$$s_n = \sum_{j=1}^{n} Z_j \quad (5)$$

where Z_j is the random event and the initial value of walk starts by 0 where $S_0 = 0$. The n in the Equation 5 determines the length of the walk and in our study the length of walk is limited with only 2 for each cases. Another variation of random walk is extending the walk length.

Figure 2. Istanbul stock market closing values for 2 years

We have tried both methods as separate signal processing approaches in our study.

4.2.2. Moving Average

Moving average is a financial analysis model which can be applied on the time series like the stock market values. There are several variations of the method like simple moving average (SMA), cumulative moving average (CMA), weighted moving average (WMA), or exponential moving average (EMA). During this study we have used SMA for the RSI calculation, which will be explained in the next subsection and moving average convergence / divergence (MACD) as a separate method (Rosenblatt, 1962).

The simplest form of moving average is calculating the mean of the given time period as demonstrated in Equation 6:

$$SMA_{t,n} = \sum_{i=t-n}^{t} \frac{P_i}{n} \tag{6}$$

where t is the current time value calculated and n is the window size on the time series and P_i is the price value on time i.

Also the SMA calculation can be followed in an iterative manner after the initial SMA value is calculated. WMA on the other hand defines some weights to the price values (P_i) depending on their distance to the current time (t):

$$WMA_{t,n} = \sum_{i=t-n}^{t} \frac{\left(n+i-t\right)P_i}{n} \tag{7}$$

The only difference between the WMA in Equation 7 and SMA in Equation 6 is the coefficients of the price values P_i where the value gets higher when the date of price gets close to the current time t.

Another variation of moving average is the exponential moving average (EMA) and the coefficients increase exponentially when the price time gets closer in the time series to the current time t:

$$EMA_{t,n} = \sum_{i=t-n}^{t} \frac{\left(1-\alpha\right)^{n+i-t} P_i}{n} \tag{8}$$

In the Equation 8 the only difference is the coefficients are getting exponential and the order of the coefficients is getting higher when the calculated price value gets closer to the calculation point on the time series. Also the value of α is the indicator of the strength of the coefficients, where higher α values indicate less importance of the older prices. Originally *EMA* doesn't deal with the number of prices handled in the window size, instead the α value determines the number of significant coefficients, but because of the implementation limitations, we have limited the number of window size by a constant variable.

Moving average convergence / divergence (MACD) is another method of analysis built on *MA* and deals with two different time length on the series, which can be named as short and long periods:

$$MACD = EMA_{short} - EMA_{long} \tag{9}$$

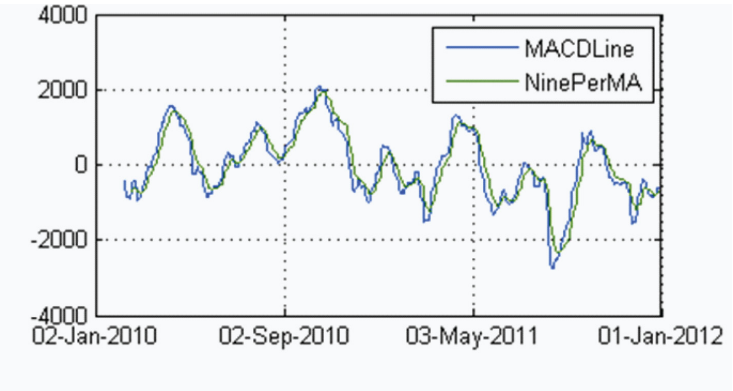
The short and long terms are built on the n parameter of EMA. For example in our application we have short as 12 days and long as 26 days and the output is shown in Figure 3.

In Figure 3, the *NinePerMA* values are the EMA_9 which is the exponential moving average for 9 days.

4.2.3. Relative Strength Index

Relative strength index (RSI) (Hecht, 1987) is a method built over the exponential moving average (EMA) which is already explained in the "moving average" sub section. For two consecutive closing values, *e.g.* $C_{now}, C_{previous} \in C$, where C is the time series of the closing values

Figure 3. MACD values of Istanbul stock market closing values



in the stock market, there are three possibilities, which are $C_{now} < C_{previous}$ or $C_{now} > C_{previous}$ or $C_{now} = C_{previous}$, for all these possibilities the RSI calculation can be demonstrated as in the Equation 10:

$$RSI = 100 - \frac{100}{1 + RS} \tag{10}$$

where RS as in Equation 11:

$$RS = \frac{EMA_{U,n}}{EMA_{D,n}} \tag{11}$$

where EMA is the exponential moving average as explained in the moving average subsection and U and D values are as in Equation 12 (see Box 1).

The RSI values after processing over the Istanbul stock market closing values are demonstrated in Figure 4.

4.2.3. Momentum and Rate of Change

Both time series analysis methods are similar to each other. Momentum aims to calculate the change on the n day interval and can be calculated as in the Equation 13 (Schalkoff, 1997):

$$momentum_n = C_{now} - C_{now-n} \tag{13}$$

Similarly rate of change can be calculated as the ratio of momentum over the interval as in the Equation 14:

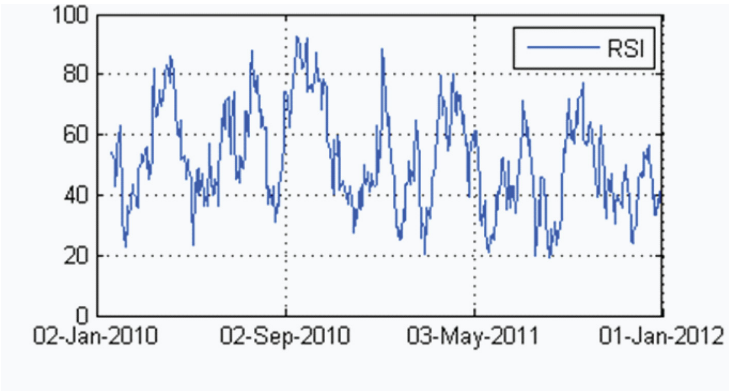
$$rate\,of\,change_n = \frac{momentum_n}{C_{now-n}} \tag{14}$$

or if the momentum value is substituted, the Equation 14 can be rewritten as in the Equation 15 (Schalkoff, 1997):

Box 1.

$$\{U, D\} = \begin{cases} U = C_{now} - C_{previous}, D = 0, if\, C_{now} > C_{previous} \\ U = 0\ , D = C_{previous} - C_{now}, if C_{now} < C_{previous} \\ \quad U = 0, D = 0, if\, C_{now} = C_{previous} \end{cases} \tag{12}$$

Figure 4. RSI values of Istanbul stock market closing values



$$rate\,of\,change_n = \frac{C_{now} - C_{now-n}}{C_{now-n}} \quad (15)$$

Momentum values are demonstrated in Figure 5 and the ROC values are demonstrated in Figure 6.

Rate of change in this case has a high alternation between -20 and +20.

4.2.4. Bollinger Band

Bollinger band processing yields a channel of upper, lower and middle Bollinger bands. The calculations can be done as in Equation 16 (Schalkoff, 1997):

$$\begin{aligned} BB_{middle,n} &= MA_n \\ BB_{upper,n} &= \sum_{i=t-n}^{t} 1, if\ MA_n > K\sigma_n \\ BB_{lower,n} &= \sum_{i=t-n}^{t} 1, if\ MA_n < K\sigma_n \end{aligned} \quad (16)$$

The σ_n values in the Equation 16 indicate the standard deviation for the given time inter-

Figure 5. Momentum values of Istanbul stock market closing values

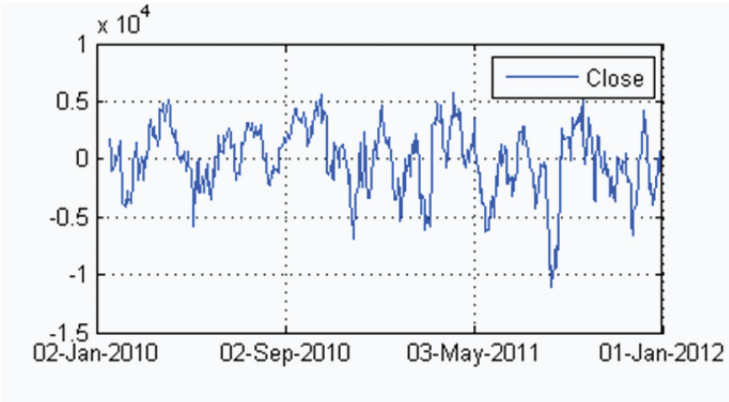
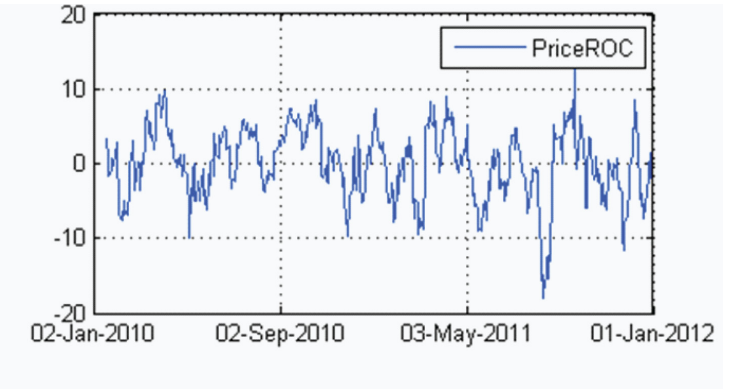


Figure 6. ROC values of Istanbul stock market closing values



val n. Also the K value is a constant, which is taken as 20 in our case.

Figure 7 demonstrates the middle Bollinger band applied over stock market closing time series.

4.2.5. Acceleration and Difference

Time series acceleration is another method of analysis. The method can be formalized as in the Equation 17 (Schalkoff, 1997):

$$Acc = diff\left(MACD_n, Histogram\right) \qquad (17)$$

The approach of acceleration is getting the difference between the moving average convergence / divergence and the histogram of the time series which is the signal itself.

The difference (or convergence / divergence) between two time series (or signals) is considered a subtraction between two series day by day.

The histogram on the other hand is the difference between MACD and the signal itself.

Figure 8 demonstrates the output of acceleration over the stock market closing values.

Figure 7. Middle BB values of Istanbul stock market closing values

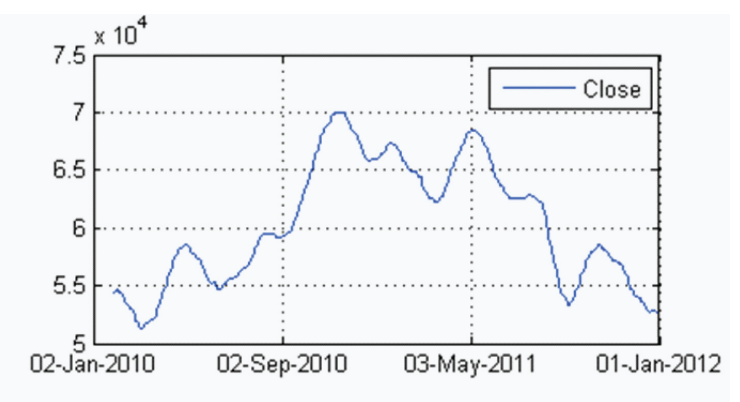


Figure 8. Acceleration applied on Istanbul Stock Market

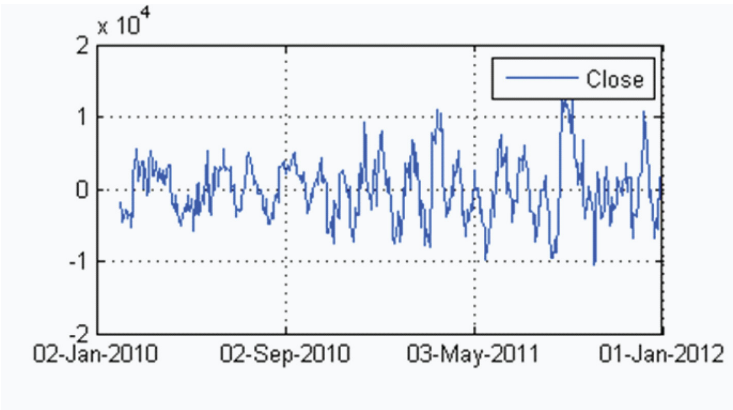


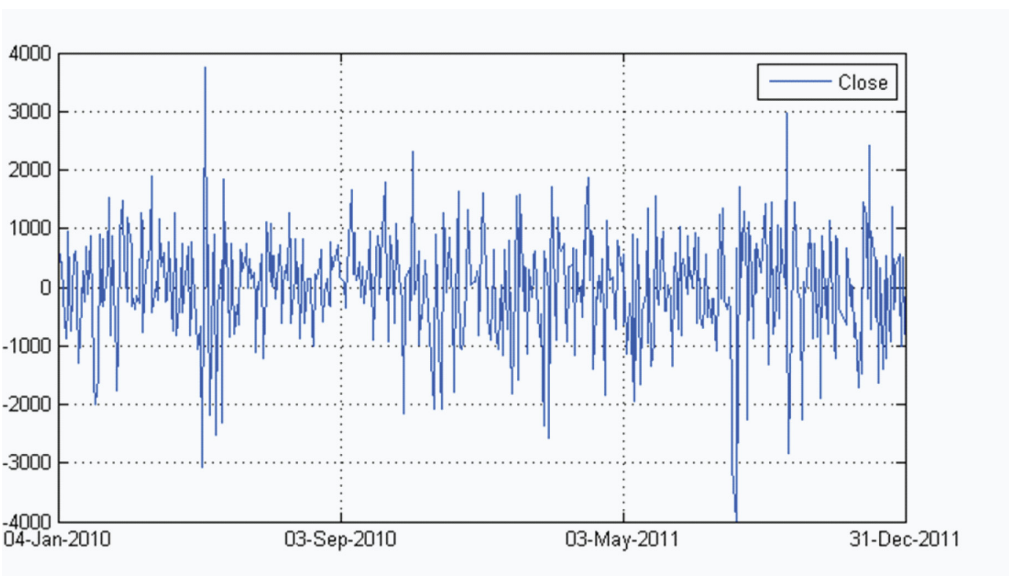
Figure 9 demonstrates the time series analysis named "difference", applied over the stock market values.

Please note that the acceleration is smoother because of the average calculation taking a part in the formulation.

4.3. Information Gain

The information gain of all the terms is calculated and ordered in descending order. Let *Attr* be the set of all attributes and E_x be the set of all training examples, *value(x,a)* with x

Figure 9. Difference applied on Istanbul Stock Market



$\in E_x$ defines the value of a specific example x or attribute $a \in Attr$, H, specifies the entropy. The information gain for an attribute $a \in Attr$ is defined as in Equation 18:

$$IG(Ex,a) = H(Ex) - \sum_{v \in v(a)} \frac{|x \in Ex \mid v(x,a)|}{|Ex|} \cdot H(x \in Ex | v(x,a)) \quad (18)$$

Also entropy in the information gain calculation can be rewritten as in Equation 19:

$$H(X) = \sum_{i=1}^{n} P(x_i) I(x_i) = \sum_{i=1}^{n} P(x_i) \log_b \left(\frac{1}{P(x_i)} \right) = \sum_{i=1}^{n} P(x_i) \log_b (P(x_i)) \quad (19)$$

4.4. Artificial Neural Network (ANN)

The purpose of ANN studies is adapting the biological neural networks into data processing. Multi-layer perceptron (MLP) is a developed version of ANN, which organizes the neurons into layers as input, output or hidden layers. Figure 10 demonstrates the approach in a data flow between layers (Rosenblatt, 1962). The ANN model has been one of the attractive tools used in geo-engineering applications due to its high performance in the modeling of non-linear multi-variate problems. Hecht-Nielsen (Hecht, 1987) and Schalkoff (1997) indicate that an ANN may be defined as a structure comprised of densely interconnected adaptive simple processing elements that are capable of performing massively parallel computations for data processing and knowledge representation.

The input layer is directly connected to the inputs which will be processed within the system. The hidden layer can hold more than one layer depending on the complexity of the system and finally the output layer holds the results. Figure 10 can be detailed if each of the layers is demonstrated by a real number of neurons. Figure 11 is a detailed version of Figure 10 with multiple neurons in each of the layers. Please note that each neuron on a layer is directly connected to all the neurons on the next layer.

Each of the connections in Figure 11, which are called a synapsis in a biological view, has a weight value, which is a coefficient to the value on the neuron which is indicated by "w_{ij}", where i and j indices are the weight value between the neurons i and j (Wasserman & Schwartz, 1988).

The aim of this study is to calculate the best possible values of the weights in Figure 11. During this calculation the back propagation algorithm is implemented. The algorithm changes the weight values on the synapses by the result achieved on the output layer.

Most of the ANN systems use the sigmoid function for the decision of firing the neuron (Mitchell, 1997). The generic formula of the sigmoid function is given in Equation 20. Mitchell uses the word "logistic function" and the "sigmoid function" synonymously. He also

Figure 10. Data flow between layers

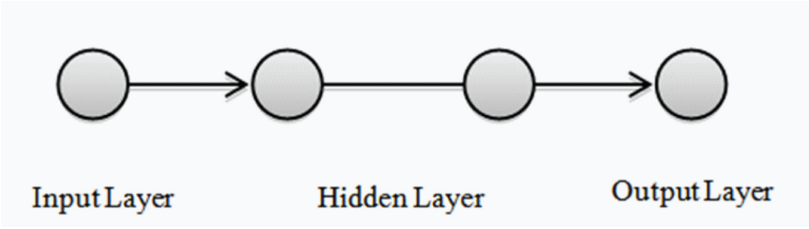
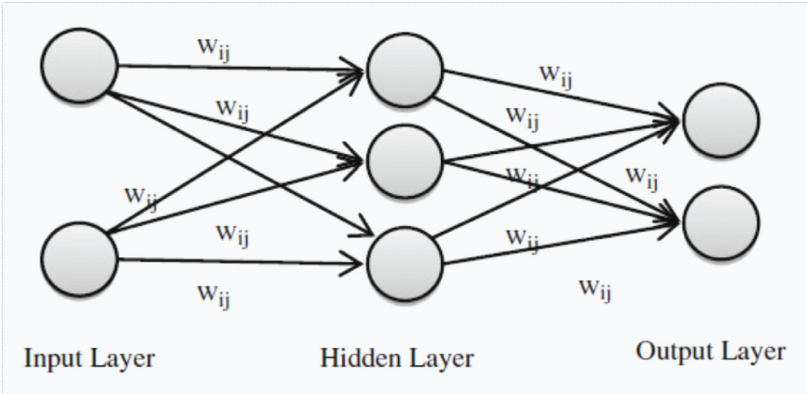


Figure 11. Generic view of MLP with multiple input and weight values of the synapses



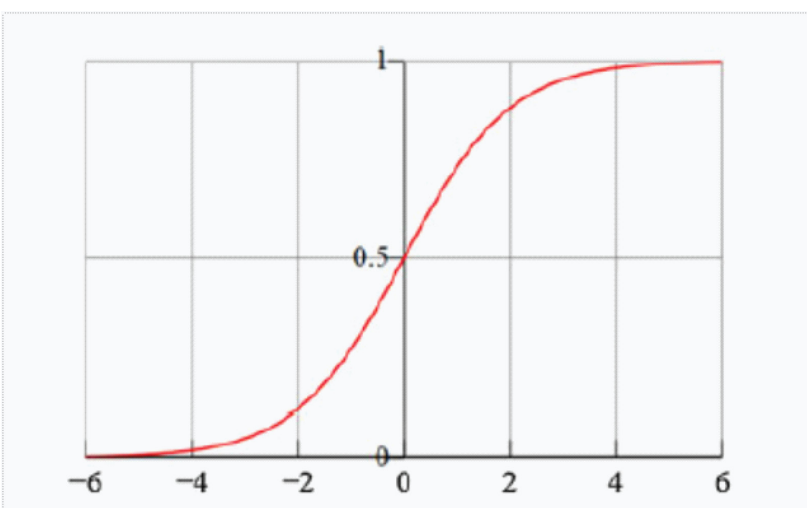
calls this function a "squashing function" and the sigmoid (a.k.a. logistic) function is used to compress the outputs of the "neurons" in multi-layer neural nets. The generic two dimensional view of the function is given in Figure 12:

$$f(x) = \frac{1}{1+e^{-x}} \tag{20}$$

Besides the above ANN information, during this study the two parameters of ANN have crucial roles which are learning rate and the momentum factor.

The learning rate of the ANN is denoted by α, which varies from 0 to 1. The learning rate is a parameter to control the amount of weight adjustment at each step of training to determine the learning rate at each step. It affects the convergence of back propagation network (BPN). A large value of α indicates the faster learning with high probability of error, while the lower learning rate indicates slower but robust learning.

Figure 12. Visualization of sigmoid function



A second crucial parameter of ANN is the momentum, which is denoted by η and varies from 0 to 1. The momentum is used to speed up the process. BPN has two major disadvantages which are larger training time and slow convergence. The momentum parameter helps to improve the convergence and the training time while reducing the oscillation of learning.

4.5. C4.5 Tree

C4.5 method (Qureshi, Mirza, & Arif, 2006) is a decision tree-based classification algorithm. The tree is built by using the information gain of each feature in the feature vector.

The algorithm starts with a training data set S, where $S= \{s_1, s_2, ..., s_n\}$ where each sample s_i has a p dimensional feature vector (FV).

For each sample s_i, $FV = \{x_{1i}, x_{2i}, ..., x_{pi}\}$ and the information gain of each values would be $IG = \{ig(x_{1i}), ig(x_{2i}), ..., ig(x_{pi})\}$.

The algorithm creates a decision tree where each node defines a decision to either side.

The highest information gain value is selected for the top most decision node and the decision criteria with the second highest gain is placed on the next level. Let ig (x_{ik}) > ig (x_{im}) for the Figure 13. The tree is constructed by following a similar approach for the next levels. Finally at the leaves, the samples are placed after the training.

At the time of testing, the features extracted from test samples are questioned via the decision nodes in the tree from root to leaves. The final leaf is accepted as the class of the test sample.

C4.5 has an advantage on other decision trees, since it uses the information gain and normalization and also it uses the pruning for the time performance.

4.6. Support Vector Machine (SVM)

The reason for applying SVM method as in Figure 14 over the dataset is to determine the boundaries between classes (25).

SVM aims to classify the samples into groups and define a boundary between the groups. SVM also tries to find out the maximum margin possibility between the groups (25):

Figure 13. C4.5 tree demonstration

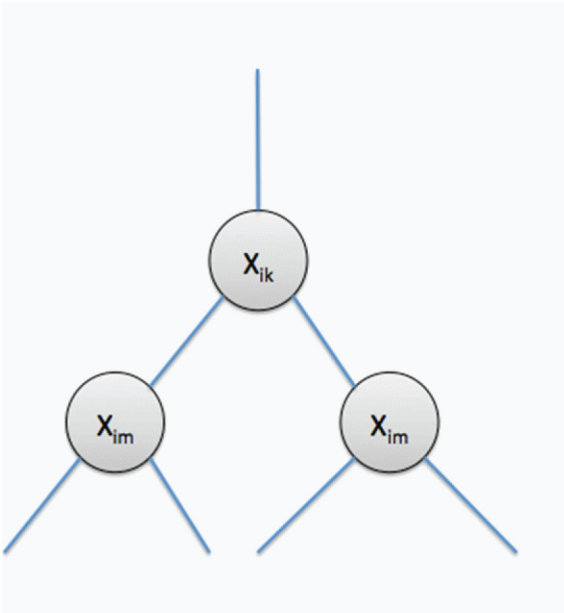
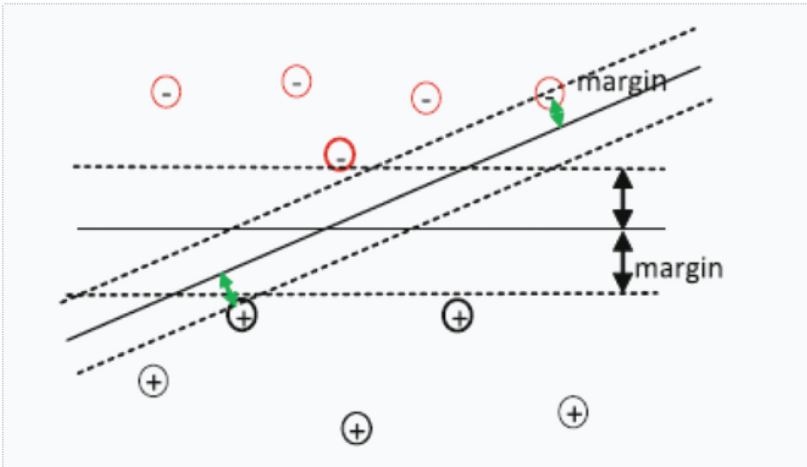


Figure 14. SVM boundary and margins



$$\omega^* = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{21}$$

The margin between the classes is symbolized by ω symbol in Equation 21 and SVM seeks to maximize the value of ω. The above formula can be rewritten as below for the linearly separable classes (26):

$$\|\omega\|^2 = \sum_{i=1}^{l} \alpha_i = \sum_{iSV_s} \alpha_i = \sum_{iSV_s} \sum_{iSV_s} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{22}$$

In the Equation 22, all the possible cases of i and j are considered. Also SVM can use a radial basis function and one of the options is the Gaussian kernel function, quoted in Equation 23 and 26:

$$K(x, x') = \exp - \left(\frac{\|x - x'\|^2}{2\sigma^2} \right) \tag{23}$$

Finally, the class is determined by the result achieved from K function.

4.7. K-Nearest Neighborhood (KNN)

The k, c-neighborhood (or k, c(x) in short) of a U-outlier x is the set of k class c instances that are nearest to x (k-nearest class c neighbors of x).

The K-NN (Wasserman & Schwartz, 1988) is explained in Figure 15. Here k is a user defined parameter. For example, k, $c_1(x)$ of an U-outliers x is the k-nearest class c_1 neighbors of x.

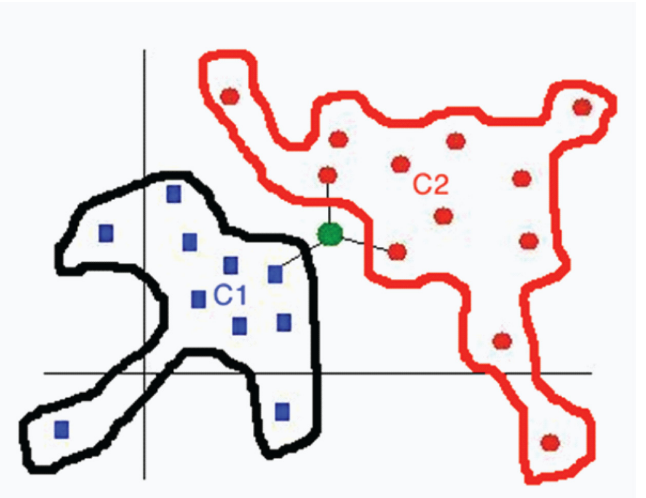
Let $\bar{D}_{Cout,q}(x)$ be the mean distance of a U-outlier x to its k-nearest U-outlier neighbors. Also, let $\bar{D}_{C,q}(x)$ be the mean distance from x to its $k, c(x)$, and let $\bar{D}_{Cmin,q}(x)$ be the minimum among all $\bar{D}_{C,q}(x)$, c ∈ {Set of existing classes}. In order words, k, c_{min} is the nearest existing class neighborhood of x. Then k-NSC of x is given in Equation 24:

$$k - NSC(x) = \frac{\bar{D}_{C\min,q}(x) - \bar{D}_{Cout,q}(x)}{\max\left(\bar{D}_{C\min,q}(x), \bar{D}_{Cout,q}(x)\right)} \tag{24}$$

4.8. Ensemble Classification

We have implemented a majority vote learning-based (MaVL or Marvel) (Masud, et al., 2011) ensemble method to combine three different

Figure 15. Visualization of K-NN



classification methods. MV can be considered as a meta classifier which works over the classifiers like KNN, C4.5 or SVM in our case.

Let $S_i \in S$ where S is the set of classifiers and let $C_i \in C$ where C is the set of classes:

$$C\left(x\right) = argmax_i \sum_{j=1}^{B} w_j I\left(S_j\left(x\right) = i\right) \tag{25}$$

where w_j is the weight of each indicator function I(•) which is added into the equation for normalization and the weights of each classifier is equal in our model.

Marvel, gets the summation for each of the classifier's vote and the sample is classified into the class with the highest vote.

4.9. Error Rate Calculation

The error rate of the system is calculated through root mean square error (RMSE). The calculation of RMSE is given in Equation 26 and 28:

$$x_{rmse} = \frac{\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}}{n} \tag{26}$$

For this study, above^x values are the results achieved from the implementation of the algorithm. The RMSE result of 0 is considered ideal and lower values close to 0 are relatively better.

By the results fetched from the output layer and the calculation of RMSE, the algorithm back propagates to the weight values of the synapses.

Also the results are interpreted by using a second error calculation method RRSE (Root Relative Squared Error) and the calculation is given in Equation 27 (Qureshi, Mirza, & Arif, 2006):

$$x_{rmse} = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{ij} - T_j\right)}{\sum_{j=1}^{n}\left(T_j - \bar{T}\right)^2}} \tag{27}$$

where P_{ij} is the value predicted for the sample case j, T_j is the target value for sample case j and $\bar{T}$ is calculated by Equation 28 (Qureshi, Mirza, & Arif, 2006):

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \quad (28)$$

The RRSE value ranges from 0 to ∞ , with 0 corresponding to ideal.

The third error calculation method is RAE (Relative Absolute Error) and the calculation is given in Equation 29 (Qureshi, Mirza, & Arif, 2006):

$$E_i = \frac{\sum_{j=1} \left| P_{(ij)} - T_j \right|}{\sum_{j=1}^{n} \left| T_j - \bar{T} \right|} \quad (29)$$

$P_{(ij)}$ is the value predicted by the individual program i for sample case j (out of n sample cases), T_j is the target value for sample case j, and $\bar{T}$ is given by the Equation 30 (Qureshi, Mirza, & Arif, 2006):

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \quad (30)$$

For a perfect fit, the numerator is equal to 0 and E_i=0 So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal.

Also the success rate of prediction and expectation can be measured as the f-measure method. The f-measure method is built on Table 1.

The calculation of f-measure can be given as in Equation 31 depending on Table 1:

$$F_{measure} = \frac{2TP}{2TP + FN + FP} \quad (31)$$

5. EXPERIMENTS

In this study the dataset is in natural language and some preprocessing for the feature extraction from the data source is required. The first approach is to apply the TF-IDF for the all terms in the data source. Unfortunately the hardware in the study environment was not qualifying the requirements for the feature extraction of all the terms in data source which is 139,434.

5.1. Dataset

We have implemented our approach and Table *2* demonstrates the features of the datasets.

The dataset in Table 2 is collected from the web site of a high-circulating newspaper in Turkey. The data is collected directly from a database so the noisy parts on the web page like ads, comments, links to other news, etc. are avoided. Another problem is the noise of HTML tags in the database entries for formatting the text of news. The data has preprocessed and all the HTML tags are removed from the news and also all punctuations and stop words are removed in the preprocessing phase.

5.2. Feature Extraction

We have implemented a feature extraction algorithm in order to extract two feature vectors.

Table 1. The f-measure method

	Predictions		
Expectations		Positive	Negative
	True	True Positive	True Negative
	False	False Positive	False Negative

Table 2. Properties of the dataset

	News
# of News	9871
Authors	6881
Texts per Author	Mean (μ): 44.05 Stddev(σ): 535.52
Average word length	~6.7

Algorithm 1 demonstrates the extraction of two vectors: one from the economic news corpus and another from the closing values of the stock market. We have limited the number of features to 300 and the Top300 function gets the topmost 300 features from the feature vector.

The V_2 feature vector is calculated easily by checking the closing value of the economic news on the date. There are some news items which are published during the time the stock market is closed like on weekends and we have considered these values as a third class besides the increase and decrease classes.

The correlation algorithms run over the two vectors V_1 and V_2 extracted via the Algorithm 1.

During the execution of the algorithm, the execution requires more memory than the available hardware, where we run the algorithms on a intel 7 CPU and 8GByte of RAM. The required memory for loading the data set is calculated in Equation 32:

$$\begin{aligned} &Memory\ Requirement = 139,434\ words \\ &\times 9,871\ news \times 6.7\ average\ word\ length \\ &\times 2\ bytes\ for\ each\ character \\ &=\sim 17GByte \end{aligned} \tag{32}$$

Also the classification algorithms require some execution space over the loaded data. In the case with four classification algorithms and a MaVL ensemble algorithm on top of these classifiers, the required memory is much higher.

As a solution we have limited the number of words with the highest occurences. The number of occurences on our implementation is 30 and a word is taken into consideration after this number of occurences. The words appearing above this threshold value is 2878 and the memory required is reduced to 700MByte which is easier to handle in the RAM.

The feature vector extraction is about 56 minutes on average for the economic news.

Algorithm 1. Feature extraction methods

```
Let E be Economy News Corpus,
Let C be Closings of Stockmarket,
For each E_i ∈ E
  For each Term_j ∈ E_i
   if(count(Term_j)>30)
   T_j ← TF-IDF of Term_j
  C_i ← closing_value(date(E_i)) ∈ C
IG_ij ← Information Gain (Term_j,E_i)
V_1 ←Top300(sort(IG))
V_2←C
```

5.3. Classification

The results of executions for the KNN algorithm can be summarized in Table 3, the execution is handled with n=3.

The success rates in Table 3 are the percentage of correctly classified instances. For example, the success rate of Random Walk with length=2 can be considered as the 37% of the instances are correctly classified to predict an increase, decrease or no change in the stock market value depending on the economic news processed.

All the nine methods are suitable for the correlation and the highest success is achieved from the Bollinger Band with 52% correctly classified news.

Table 4 summarizes the results gathered from the decision tree algorithm. Again the Bollinger band analysis method is leading with a success rate higher than 53%.

SVM classification method has the highest success rate among all the classification methods. The success rate as demonstrated in Table 5 is almost 54% which is slightly higher than the C4.5 decision tree approach.

Table 3. Error and success rates of KNN, n=3

	f-Measure Average	RMSE	RAE	Correctly Classified
Random Walk	0.508	0.4398	0.9924	51.58%
Bollinger Band	0.529	0.4383	0.9883	52.92%
Moving Average	0.507	0.4375	0.9937	52.07%
Momentum	0.504	0.4436	0.9974	50.37%
Difference	0.505	0.444	1.0009	50.45%
RSI	0.501	0.4404	0.9930	50.70%
ROC	0.505	0.4422	0.9933	50.86%
Periodic Average	0.508	0.443	0.9960	50.77%
Random Walk Length=2	0.297	0.585	0.9627	37.67%

Table 4. Error and success rates of C4.5 decision tree

	f-Measure Average	RMSE	RAE	Correctly Classified
Random Walk	0.499	0.442	1.0010	50.89%
Bollinger Band	0.526	0.4318	0.9753	53.71%
MACD	0.504	0.442	0.9929	51.85%
Momentum	0.509	0.441	0.9906	51.07%
Difference	0.511	0.4427	0.9928	51.51%
RSI	0.496	0.4444	0.9982	50.35%
ROC	0.504	0.4419	0.9917	50.78%
Periodic Average	0.517	0.4416	0.3308	48.25%
Random Walk Length=2	0.499	0.4442	1.0010	50.89%

Table 5. Error and success rates of SVM

	f-Measure Average	RMSE	RAE	Correctly Classified
Random Walk	0.442	0.4225	0.9828	53.11%
Bollinger Band	0.452	0.4207	0.9793	53.82%
MACD	0.437	0.4227	0.9811	53.23%
Momentum	0.482	0.4295	0.9954	50.44%
Difference	0.455	0.4282	0.9936	50.93%
RSI	0.47	0.4271	0.99	51.37%
ROC	0.485	0.4313	99.95	49.77%
Periodic Average	0.476	0.4295	0.3326	50.45%
Random Walk Length=2	0.442	0.4323	0.9822	53.09%

Table 6. Error and success rates of MaVL

	f-Measure Average	RMSE	RAE	Correctly Classified
Random Walk	0.497	0.4182	0.9921	52.37%
Bollinger Band	0.504	0.4141	0.3257	53.49%
MACD	0.491	0.4174	0.9892	52.52%
Momentum	0.497	0.4209	1.0297	50.81%
Difference	0.491	0.4205	0.9957	50.7%
RSI	0.491	0.4202	0.9938	50.87%
ROC	0.496	0.4209	0.9948	50.24%
Periodic Average	0.505	0.4202	0.3322	51.46%
Random Walk Length=2	0.497	0.4184	0.9921	52.37%

Similar to the other three classification methods, the ensemble classification has a higher impact on the Bollinger band analysis. On the other hand, the success rate of majority voting is sometimes increasing and sometimes decreasing, depending on the time series analysis method.

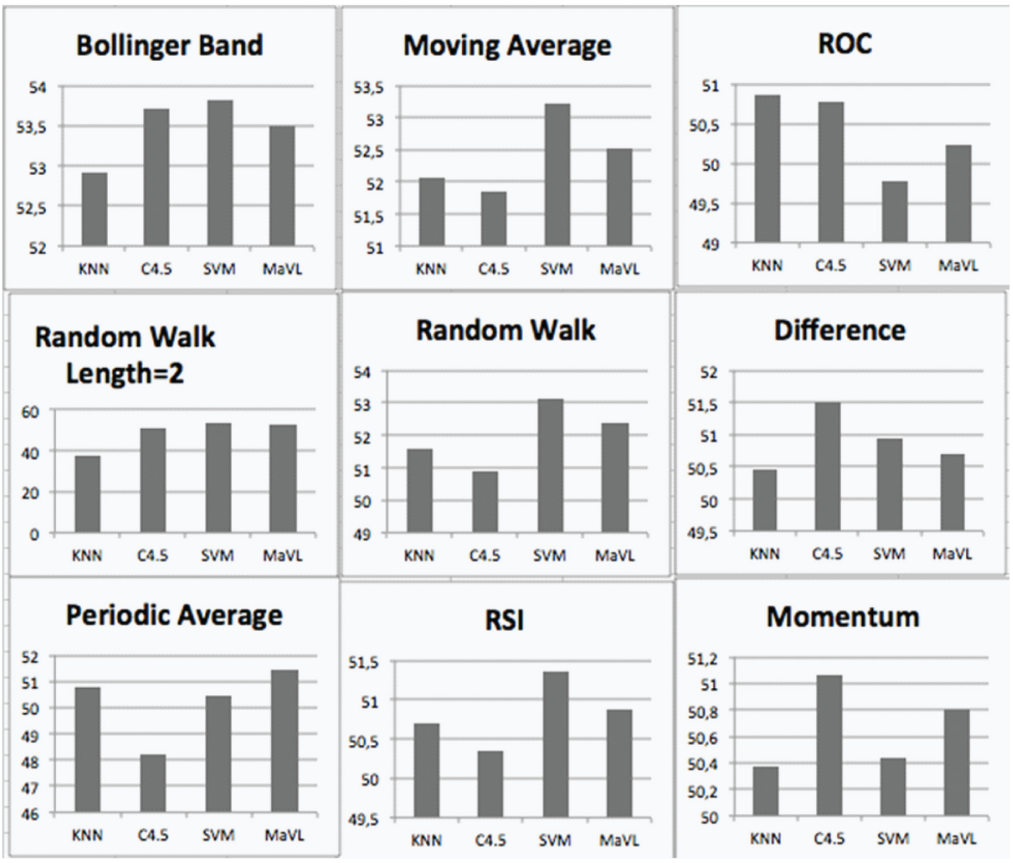
The success rates can be summarized as in Figure 16. All nine time series analysis methods are demonstrated.

The value of success is highly related with the market structure so the success rate here should not be understood as the success rate of the methodology or the classifier. The success rate in the table is the correlation between economic news and the stock market closing values.

6. CONCLUSION

In this paper, we have studied the correlation between a stock market and economic news for the first time for a Turkey case. The correlation is strongly related tothe market structure and how it is speculative depending on the economic news. The results show that for the best classi-

Figure 16. Summarization of the success rates from time series analysis



fier we have applied, there is a 53% correlation between these two datasets. This study also for the first time takes care of the effect of time series analysis and the classification method together. Nine different time series analysis and four different classification are tested on the datasets and results are demonstrated. From the results we can conclude that the Bollinger band and SVM methods are the most successful methods among the methods studied in this paper.

During this study, we have also dealt with text mining over big data and feature reduction because of the physical limitation and time series analysis approach over the stock market.

We believe the results would be beneficial for further research on the finance and economy fields as well as the data mining field. As a future work, more classifiers can be added or an ensemble approach can be applied over the classification phase and also some new feature extraction methods can be applied like using part of speech tagging and so on.

REFERENCES

Braun, H., & Chandler, J. S. (1987). Predicting stock market behavior through rule induction: an application of the learning-fromexample approach. *Decision Sciences*, *18*(3), 415–429. doi:10.1111/j.1540-5915.1987.tb01533.x

Fung, G. P., Yu, J. X., & Lam, W. (2002). News sensitive stock trend prediction. *Lecture Notes in Computer Science*, *233*, 481–493.

Halgamuge, S. Y, Z., & Hsu, A. (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. Advances in Neural Networks - ISNN 2007 (Lecture Notes in Computer Science). 4493, pp. 1087-1096. Springer-Verlag Heidelberg.

Hecht, N. R. (1987). Kolmogorov's mapping neural network existence theorem. *IEEE First Annual International Conference on Neural Networks, 3*, pp. 11-14. San Diego, USA.

Lu, H.-M., Huang, N. W., Zhang, Z., & Chen, T.-J. (2009). Identifying Firm-Specific Risk Statements in News Articles. *PAISI '09 Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics* (pp. 42-53). Springer-Verlag Berlin, Heidelberg.

Mahajan, A., Dey, L., & Haque, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market. *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on, 1*, pp. 423-426.

Mahmoud, S., & El-Melegy, M. T. (2004). Evaluation of diversity measures for multiple classifier fusion by majority voting. *Electrical, Electronic and Computer Engineering, 2004. ICEEC '04. 2004 International Conference on*, (pp. 169- 172).

Masud, M. M., Al-Khateeb, T. M., Khan, L., Aggarwal, C., Gao, J., Han, J., & Thuraisingham, B. (2011). Detecting Recurring and Novel Classes in Concept-Drifting Data Streams. *ICDM '11 Proceedings of the 2011 IEEE 11th International Conference on Data Mining* (pp. 1176-1181). Washington, DC, USA: IEEE Computer Society.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

Mittermayer, M.-A. (2004). Forecasting intraday stock price trends with text mining techniques. *HICSS '04 Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04). 3*, pp. 64-73. IEEE Computer Society Washington, DC, USA.

Mittermayer, M.-A., & Knolmayer, G. F. (2006). NewsCATS: A News Categorization and Trading System. *ICDM '06: Proceedings of the Sixth International Conference on Data Mining* (pp. 1002-1007). IEEE Computer Society.

Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, 4*, pp. 256- 260.

Qureshi, S., Mirza, S. M., & Arif, M. H. (2006). Fitness function evaluation for image reconstruction using binary genetic algorithm for paralel ray transmission tomography. *Emerging Technologies, 2006. ICET '06. International Conference on*, (pp. 196- 201). Islamabad, Pakistan.

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADMIRAL: A Data Mining Based Financial Trading System. *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on* (pp. 720 - 725). doi: 10.1109/CIDM.2007.368947.

Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Washington: Spartan Books.

Schalkoff, R. J. (1997). *Artificial neural network*. New York: McGraw Hill.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin Text system. [TOIS]. *ACM Transactions on Information Systems*, *27*(2), 1–19. doi:10.1145/1462198.1462204

Soni, A., Eck, V., Jan, N., & Uzay, K. (2007). Prediction of stock price movements based on concept map information. *IEEE Symposium on Computational Intelligence in Multicriteria Decision Making*, (pp. 205- 211). Honolulu, HI.

Tan, F. H. (n.d.). Interpreting News Flashes for Automatic Stock Price Movement Prediction. *Erasmus University Rotterdam.*

Trappe, W., & Washington, L. C. (2006). *Introduction to Cryptography with Coding Theory*. Pearson Prentice Hall.

Wasserman, P. D., & Schwartz, T. (1988). Neural networks II. What are they and why is everybody so interested in them now? *IEEE Expert*, *3*(1), 10–15. doi:10.1109/64.2091

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. *SMC98 Conference Proceedings 1998 IEEE International Conference on Systems Man and Cybernetics Cat No98CH36218. 3*, pp. 2720-2725. Ieee.

Yahia, M. E., & Ibrahim, B. (2003). K-nearest neighbor and C4.5 algorithms as data mining methods: advantages and difficulties. *Computer Systems and Applications, 2003. Book of Abstracts. ACS/IEEE International Conference on*, (p. 103). Tunis, Tunisia.