I.J. Intelligent Systems and Applications, **2017, 7, 22-30**
Published Online July 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.07.03

Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis

Ayman E. Khedr
Faculty of Computers and Information Technology, Future University in Egypt,
Information Systems Department, Cairo, Egypt
E-mail: Ayman.khedr@fue.edu.eg

S.E.Salama
Faculty of Computers and Information, Helwan University,
Information Systems Department Cairo, Egypt
Faculty of Computers and Information Technology, King Abduaziz University, Jeddah, Kingdom of Saudi
E-mail: Chaimaa_Salama@yahoo.com

Nagwa Yaseen
Faculty of Computers and Information, Helwan University,
Information Systems Department Cairo, Egypt
E-mail: Nagwa.yaseennm@gmail.com

***Abstract*—**Stock market prediction has become an attractive investigation topic due to its important role in economy and beneficial offers. There is an imminent need to uncover the stock market future behavior in order to avoid investment risks. The large amount of data generated by the stock market is considered a treasure of knowledge for investors. This study aims at constructing an effective model to predict stock market future trends with small error ratio and improve the accuracy of prediction. This prediction model is based on sentiment analysis of financial news and historical stock market prices. This model provides better accuracy results than all previous studies by considering multiple types of news related to market and company with historical stock prices. A dataset containing stock prices from three companies is used. The first step is to analyze news sentiment to get the text polarity using naïve Bayes algorithm. This step achieved prediction accuracy results ranging from 72.73% to 86.21%. The second step combines news polarities and historical stock prices together to predict future stock prices. This improved the prediction accuracy up to 89.80%.

***Index Terms*—**Data Mining, Stock Market, sentiment analysis, Text Mining, news sentiment analysis.

I. Introduction

Stock market decision making is a very difficult and important task due to the complex behavior and the unstable nature of the stock market. There is an important need to explore the enormous amount of valuable data generated by stock market. All investors usually have the imminent need of finding a better way to predict the future behavior of stock prices, this will help in determining the best time to buy or sell stocks in order to achieve the best profit on their investments. Trading in stock market can be done physically or electronically. When an investor buys a company stock, this mean that this investor becomes an owner of the company according to the ownership percentage of this company's shares. This give the stockholders rights on the company's dividends [1]. Financial data of stock market is of complex nature, which makes it difficult to predict or forecast the stock market behavior. Data mining can be used to analyze the huge and complex amount of financial data, which leads to better results in predicting the stock market behavior. Using data mining techniques to analyze stock market is a rich field of research, because of its importance in economics, as better prices lead to an increase in countries' income. Data mining tasks are divided into two major categories; descriptive and predictive tasks [2], [3]. In our study we consider the predictive tasks. Classification analysis is used to predict the stock market behavior. We use Naïve Bayes and K-NN algorithms to build our model.

The prediction of stock market helps investors in their investment decisions, by providing them strong insights about stock market behavior to avoid investment risks. It was found that news has an influence on the stock price behavior [4]. Stock market prediction based on news mining is an attractive field of research, and has a lot of challenges because of the unstructured nature of news. News mining can be defined as the process of extracting hidden, useful and potentially unknown patterns from news data to obtain knowledge. Text mining is a technique used to handle the unstructured data. Text mining also known in data mining as the step of Knowledge Discovery in Text (KDT). Walter et al. [4] investigate the relation between financial news and stock market volatility using granger causality. The study

reveals that there is a relation between news sentiment and stock prices changes.

Sentiment analysis is the process of determining people's attitudes, opinions, evaluations, appraisals and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [5]. Sentiment analysis considered a particular branch of data mining that classifies textual data into positive, negative and neutral sentiments [28].

Zubair et al.[6] analyze the correlations between Reuters news sentiment and S&P500 index for five years data. This is done using Harvard general inquirer to obtain positive or negative sentiment, then kalman filter tool is used for smoothing estimation and noise reduction. The results demonstrate that there is a strong correlation between S&P500 index and negative economic sentiment time series. Text preprocessing [7], [8] is a vital and important task in text mining, NLP and information retrieval. It is used for preparing unstructured data for knowledge extraction. There are many different tasks for text preprocessing; tokenization, stop-word-removal and stemming are among the most common techniques. Tokenization is the process of splitting the text into a stream of words called tokens. Tokenization have an importance in linguistics and computer science fields and considered a part of lexical analysis. Identifying the meaningful keywords is the main goal of using tokenization. Stop-word-removal is the process of removing the frequently repeated words that does not have any significant meaning in the document such as the, and, are, this…etc. Stemming aims at return the variation of the word into common representation by removing suffixes [7].

In this paper, the proposed approach uses sentiment analysis for financial news, along with features extracted from historical stock prices to predict the future behavior of stock market. The prediction model uses naïve Bayes and K-NN algorithms. This is done by considering different types of news related to companies, markets and financial reports. Also, different techniques for numeric data preprocessing as well as text analysis for handling the unstructured news data. The competitive advantage of stock market trend prediction achieved by data mining and sentiment analysis includes maximization of profit, minimizing costs and risks along with improving the investor's awareness of stock market that leads to accurate investment decisions.

II. Related Work

Several approaches for predicting stock market behavior and prices trend have been studied in literature. Some of these studies focus on improving the accuracy of prediction based on sentiment analysis of news or tweets along with stock prices such as [9]. Others focus on price prediction with different time frames such as [10]. Moreover, different research approaches proved that there is a strong correlations between financial news and stock prices changes like [4], [6]. Finally, research studies were conducted to improve the prediction accuracy such as [11], [12].

All previous studies have a challenge because of the complexity of dealing with unstructured data. All approaches are based on text mining techniques to predict stock market trend, some of them depend on textual information compared with only closing prices and others depend on textual information and stock prices charts screen tickers such as [6].

A. Studies Relaying On Social Media Information Analysis

L.I. Bing et al. [13] proposed an algorithm to predict the stock price movement with accuracy up to 76.12% by analyzing public social media information represented in tweets data. Bing adopted a model to analyze public tweets and hourly stock prices trend. NLP techniques have been used along with data mining techniques to discover relationship patterns between public sentiment and numeric stock prices. This study investigates whether there is an internal association in the multilayer hierarchical structures, and found that there is a relation between internal layers and the top layer of unstructured data. This study considers only daily closing values for historical stock prices. Y. E. Cakra [14] proposed a model to predict Indonesian stock market based on tweets sentiment analysis. The model has three objectives: price fluctuation prediction, margin percentage and stock price. Five supervised classification algorithms have been used in tweets prediction: support vector machine, naïve bayes, decision tree, random forest and neural network. This study proved that random forest and naïve bayes classifiers outperformed the other used algorithms with accuracy 60.39% and 56.50% respectively. Also, linear regression performs well on prices prediction with 67.73% accuracy. The limitation of this study is that the prediction model is constructed based only on the prices of five previous days.

Hana and Hasan [9] used hourly stock news with breaking tweets along with one hour stock prices charts to predict if hourly stock price direction will increase or decrease. This study investigates whether the information in news article with breaking tweets volume indicates statistical significant boost in hourly directional prediction. The research results demonstrated that logistic regression with 1-gram keyword performed well in directional prediction, also using extracted document level sentiment features does not have a statistical significant in boosting hourly directional prediction, but this study depends on only breaking news for hourly prediction.

B. Studies Relaying On News Analysis

Patric et al. [10] used several integrating text mining methods for sentiment analysis in financial markets by integrating word association and lexical resources to analyze stock market news reports. The study analyzes German language using sentiWS tool for sentiment analysis on different levels. The stock prices screens are compared to sentiment measures model to get investor's recommendation for one week to help them avoid

investment risks. Shynkevichl et al.[15] used multiple kernel learning (MKL) methods to investigate using two categories of news, articles related to sub-industry and articles related to a target stock.

The research investigates if these two categories will enhance the prediction of stock trend accuracy depending on news data and historical stock prices data. Historical stock prices used in Shynkevichl's study are open and close attributes. This study reveals that using different categories of news will enhance the accuracy of prediction up to 79.59 % when polynomial kernels are used on news categories. The study also proved that using support vector machine and k-NN achieve worse prediction accuracy. In [16] association rule mining has been used to uncover stock market patterns and generate rules to predict the stock price through helping the investors in the investment decisions. The prediction has been done through giving investors clear insight to decide whether to buy, sell or hold shares. Association rule mining used important six trading technical indicators to generate rules. Naive Bayes algorithm has been used to predict the class label for investor like sell, buy and hold for each stock. This is done through considering the effects of all technical indicator values and calculate the technical indicator that has the highest probability. The limitation of this research is using the closing price only without using the textual financial information, which is insufficient to provide information about event extraction financial news. Ho'ang and Phayung [11] proposed a model to predict stock price trend using Vietnam stock index prices data and news information of news publications . In this study, support vector machine algorithm is combined with linear SVM. The results of Hoang's model demonstrate that the accuracy of prediction is improved up to 75%. This study also used the closing prices of the index prices only to predict the trend.

Jageshwer and Shagufta [12] analyzed the impact of financial news on the stock market prices prediction and daily changes in the index movements. The focus of this study is to improve the accuracy of the prediction by combining technical analysis and the rule based classifier. The prediction model depends on the financial news and monthly average for daily stock price. Ruchi and Gandhi [17] presented a model to predict the stock trends by analyzing non-quantifiable information that is presented in news articles. NLP methodology is built in this model using senti-wordnet 0.3 along with the statistical parameter based module. The model used stock intrinsic values of open and close to output the sentence polarity and the behavior to be either positive or negative. The obtained behavior is based on a statistical parameter, however this study can be improved using other attributes that can affect the stock prices directly along with the data mining prediction algorithms.

Sadi et al.[18] investigated the correlation between the economic news and time series analysis methods over the charts of the stock market closing values. Ten methods have been applied for time series analysis along with using SVM and KNN classifiers. Y.Kim et al.[19] explored the stock market trend prediction using opinion mining analysis for the economic news. Kim's study assumed that there is a strong relation between news and stock prices changes to be either positive or negative changes. This model is built using NLP, news sentiment and opinion mining based sentimental dictionary. This study achieved an accuracy of prediction ranging from 60 % to 65%. S.Abdullah et al.[20] analyzed Bangladesh stock market using text mining and NLP techniques to extract fundamental information from textual data. This study used the information parser algorithm and Apache OpenNLP which is a java based machine learning toolkit for natural language processing to analyze textual data related to the stock market. This study considered the different fundamental factors includes, EPS, P/E ratio, beta, correlation and standard deviation along with price trend from historical data to compare it to the extracted fundamental information. The aim of this study was to help investors make their investment decisions for buy or sell signals.

The previous conducted researches are based on textual data analysis, and they achieved accuracies that do not exceed a range of 75% to 80% for stock trend prediction. In news polarities, the predictions accuracy range does not exceed 76%. The proposed study in this paper, aims at minimizing losses by achieving high accuracy in prediction based on sentiment and historical numeric data analysis.

The discussed pervious researches differ in prediction horizon, some of them predict prices fluctuation for 5 to 20 minutes, hourly and daily after news releases. Among the previous researches goals is to obtain investors recommendation such as [10], others is to predict only news polarities compared with actual trend from historical data.

Attempts to predict the stock market along the history is not just limited to data mining models, there are a lot of studies designed to predict the stock market using neural networks and artificial intelligence such as [29],[30].

In this study, we aim to construct a model to predict news sentiment using NLP techniques and then predict the future stock price trend using data mining techniques. The proposed study presents a new approach with improved prediction accuracy to avoid the big losses and risks of investment and maximizes the stock market profits thus avoids the Economic crises.

III. Proposed Model

The proposed model helps investors avoid risks and financial crises when making investment decisions.

The goal of the proposed model is to predict the stock market behavior, whether it is falling or raising. The proposed architecture combines the analysis of the stock market news and the historical prices together, in order to boost the classification accuracy of the stock market behavior. The study proposes performing text analysis on stock market news to determine the polarity of the news articles. Moreover, stock market historical prices opening, high, low and closing prices (OHLC) are analyzed to

predict the future trends.

Open price is the open value of the stock in the current day, high and low prices are the highest and lowest values of the stock during a day respectively, and close price is the closing value of the stock for the current day.

To achieve the required target, the following tasks are applied as shown in Fig. 1.

- Sentiment analysis of stock news articles is performed in the sentiment analysis component. In this step, preprocessing techniques are followed by Naive Bayes algorithm to determine the sentiment for each news article or company's financial reports.
  In the numeric data analysis component, preprocessing step is performed on numeric stocks data. The output of this process are two separate feature vectors, one containing labeled news articles to be either 'positive' or 'negative', and the other containing numeric stock data.
- The two feature vectors are combined based on the date of each stock's news and numeric data. This forms one concatenated feature vector containing numeric and sentiment data for each stock. Finally, K -Nearest Neighbor algorithm is used to predict the future stock price behavior, either 'raise' or 'fall'.
  Details about the architecture are presented in the following sections.

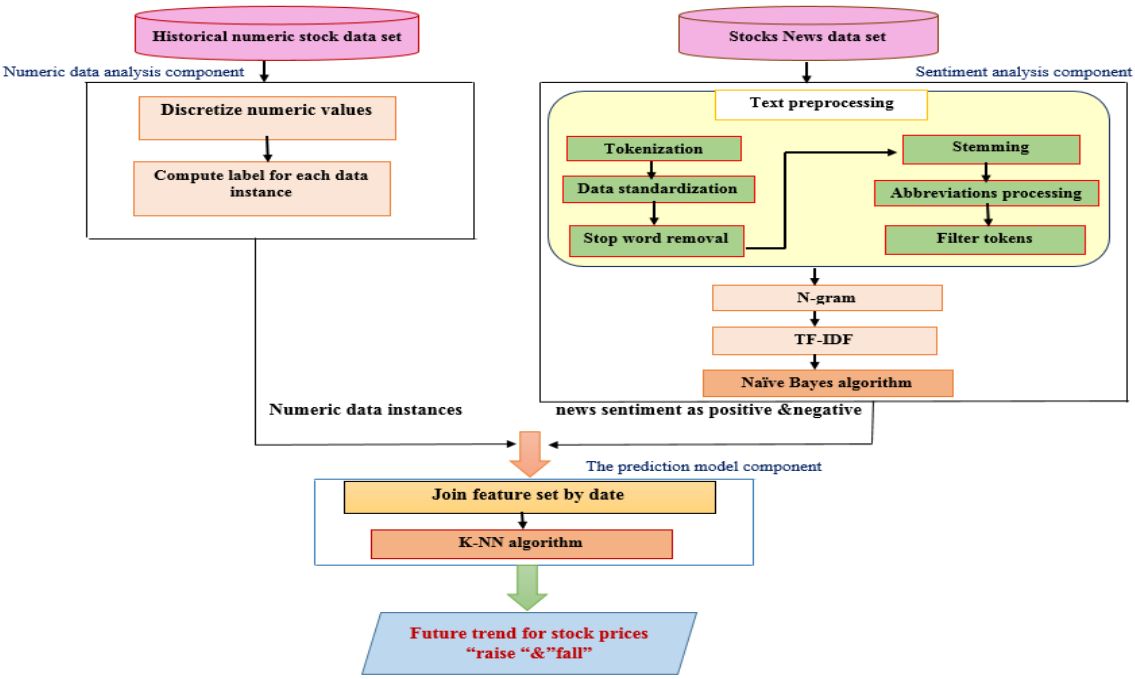

Fig.1. The Proposed Model For Stock Trend Prediction

A. Data Description

In this study data of three companies traded in NASDAQ are used. NASDAQ is "the largest stock-based equity securities market in the United States"[21],[22]. These three companies are yahoo Inc, Microsoft Corporation MSFT and Facebook Inc (FB Inc). Two data sets are collected, news and numeric data. For news, data is collected from different data sources such as nasdaq.com, Reuters, wall street journal, marketwatch.com, zacks.com, yahoo finance, Google finance, ecomomics.com. Each day news is posted about the market and about companies, some of these are news article and others are considered financial reports for the companies. We collected three news articles per day. The considered news in this study are news about company dividends or stocks dividends, news about outstanding shares changing, company splitting, merge of companies stocks by financial experts and financial reports.

For numeric data, we considered the following attributes opening, high, low and closing prices (OHLC) as these have direct effect on the prediction of the future stock prices trend.

B. Proposed Model Component Description

This section describes each component of the proposed model. We begin by explaining the first component:

1) Sentiment Analysis Component:

In this component, the analysis of stock news data is performed as follows: For news data, the objective is to classify news to be either positive or negative sentiments. To achieve this, data preprocessing is performed on news text, followed by news classification using Naïve Bayes algorithm. The following section describes the details of the proposed steps.

- Text preprocessing

There are a several preprocessing steps are performed as follow:

Tokenization: Each news article or financial report document is split into meaningful words called tokens.

Data standardization: Using data standardization techniques for data consistency, this is done by transforming all words in articles and reports about companies in a document into lower case.

Stop-word-removal: Words that do not have a significant meaning in the documents such as: the, a, of…etc. are removed to reduce the number of features and improve the performance.

Stemming: Porter stemmer is applied on the data to return each word to its stem and remove suffixes such as (-Ed,-ing,-ion…etc.) to reduce the complexity in the document and minimize the processing time which improve the model performance.

Abbreviation processing: Creating a list of abbreviations such as "U.S" for United States, "FB Inc." for Facebook income …etc. then replace the abbreviations.

Filter tokens: Words that consist of two or less characters are filtered.

In document representation we use a word vector representation that represents each word as a real vector to reduce the complexity of dealing with text data.

- N-Gram:

After data preprocessing, N-Gram features extraction is performed. N-gram is a series of tokens or words of length n and is used in many text mining and NLP tasks. The proposed model creates n-gram for stocks news documents to extract keyword features from news corpus.

Setting n=2 means that a sequence of two-words for each document is generated. This step raises the accuracy of the classifier because of the achieved information or features from two sequence of word combinations. N-gram based extraction is proved to have a robust performance in extracting features from text due to different aspects. First, the automatic capturing for the most frequent roots in stock market data. Second, the good representation that is provided by n-gram, does not require using a specific dictionary. Third, its tolerance for deformation and spelling errors[23].

- TF-Idf:

(Term frequency –inverse document frequency) a feature weighting method which is used to reveal the importance of the words in the document or a collection of corpus. The proposed model uses TF-Idf to determine the value of each word in a document through the ratio of Idf in a specific document to all documents that have the word appearance. Words that have high value imply that there is a strong relation with the document that it appears in. Equation (1) represents tf-idf calculation [24],[25] where tf is the term frequency for term t in document d, that gives weight for the term based on the term occurrence in a document. Inverse document frequency Idf adjusts the weight of processing a token for an item based upon the number of items that contain the term in the existing database[26]. This is because a term that occurs in a few documents is likely to be a better discriminator than a term that appears in most or all documents. According to equation 1, Tf-Idf is calculated where nt is the number of documents that have term t and n is the total number of documents to be analyze.

$$tf - idf = \frac{n_t}{n} * \log_2(n/n_t) \quad (1)$$

- Naïve Bayesian classifiers:

Naïve Bayes classifier is used to classify the stock news to be either positive or negative sentiment based on TF-idf values. The Naïve Bayes algorithm assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional Independence. Naïve Bayes classifier has been used to predict the polarity of each document because of its simplicity and speed in text classification. It assigns each document into the positive or negative class:

$$PNB(\mathrm{c}|\mathrm{d}) = \frac{P(c) \times (\prod_{i=1}^{m} P(f_i|c)^{ni(d)})}{P(d)} \quad (2)$$

Equation (2) represents the calculation of naïve Bayes classifier where fi is the feature that appears in the document and n_i (d) are the number of occurrences of the features in the document. Naïve Bayes classifier determines the polarity of the news to be either positive or negative with reasonable accuracy of the model for three different companies as follow yahoo, Msft, Fb Inc.

2) Stocks Numeric Data Preprocessing Component

For numeric stock prices, the features used are open, high, low and close. The first step, discretizes numeric values by converting all numeric attributes into discrete values to be positive, negative or equal [19]. This is performed by comparing all numeric values for each attribute with previous closing price value. After comparison, if the value of the attribute is greater than the previous closing price value, then this attribute value will be replaced with "positive", if the attribute value is less than the previous closing price value it will be replaced with "negative".

Finally, if the value of attribute is equal to the value of closing price then it is replaced with "equal". Briefly this means that all four attributes values will be labelled with "positive", "negative "and "equal" based on the comparative with the previous day closing price. The second step, computes a label for each data instance, the trend 'raise' or 'fall' is calculated based on the difference between the closing price of the current day and the previous day. The output of this stage is used as input to the prediction model.

3) The Prediction Model component

The proposed prediction model combines news data with numerical stock prices to investigate the influence of news releases and numeric data on stock raising and falling movements. The prediction model component consists of two steps, they are described in the following

sections.

- Joining Feature Set by Date

Both feature vectors from the previous step are augmented together. News sentiment and numeric data are joined by stock date; resulting in the following features: news sentiment, open, high, low and closing prices for each day. It is important to note that we rely on multiple news for each day instead of one, which will provide more information about the stock prices behavior.

- Stock Prediction (KNN classifier)

The final step is to predict the stock class based on the collected feature set. Data is split into training and testing sets, and K-NN classifier is used to predict the stock trend fall or raise. KNN classifier is a method for classifying objects based on the closest training examples in the feature space. The class label is assigned the same class as the nearest K instances in the training set. KNN is a type of lazy learner strategy that delays the process of applying the model on training data only if it is necessary to classify test data. KNN classifier is considered a flexible and simple classification technique where information about the training data distribution is available [3], [27].

IV. Experiment and Evaluation

This section describes the experiment results performed to predict the stock market behavior using data mining and news sentiment analysis. The experiment consists of two phases, the first phase describes the results of news sentiment analysis component that classifies news into positive or negative. The second section describes the result of the prediction model for the stock market behavior to be either positive or negative. In both phases of the experiment three company's data have been used. These companies are yahoo inc, Google inc and Facebook inc. In the sentiment analysis phase, the news data is divided into training and testing sets. The training set is used to learn the model and the testing set to validate it. The training data contains the news content with its sentiment as positive or negative to learn the model. The testing data contains the news content and the model should be able to correctly classify the testing data instances into positive or negative news. The following sections describes the results of the experiments.

A. Results of News Sentiment Analysis Component

The results of news sentiment analysis model are presented in table 1. The results show that accuracy achieved for sentiment analysis model is up to 86.21% for yahoo Inc Company, 82.76% accuracy for FB Inc and 72.73% for Msft company. This high achieved accuracy of our model compared to previous researches is due to using naïve Bayes algorithm along with NLP techniques and Tf-idf. Naive Bayes algorithm gives good performance, high accuracy and performed well with textual data. As shown in the obtained results we achieved higher accuracy for news polarities than other previous studies. The lower accuracy results of Msft Inc are due to lower number of data instances compared to news collected for Yahoo and FB.

Fig. 2 shows samples of actual and predicted sentiment polarities as positive or negative for news data. The representation of the feature weighting method Tf-idf that reveals the importance of the word in the document or a collection of corpus along with n-gram that creates two n-gram for stocks news documents and extracted keyword features from news corpus as shown in fig. 2.

Sentiment	predicted sentiment	fell	fell-expect	fell-unexpectedly	earn	earn-season	earn-growth	close	close-point	board-director	open	open-session
positive	positive	0.078	0.148	0	0	0	0	0.032	0.021	0	0.057	0
negative	negative	0	0	0	0	0	0	0.029	0.013	0	0.026	0.055
positive	positive	0	0	0	0.096	0	0.035	0	0	0	0	0
negative	negative	0.044	0	0	0	0	0	0	0	0	0.032	0
positive	negative	0	0	0	0.125	0.050	0	0	0	0.042	0	0
negative	negative	0	0	0	0	0	0	0.106	0.020	0.061	0.014	0.030
negative	positive	0.075	0.142	0.120	0.044	0.142	0	0.031	0	0	0.054	0
positive	positive	0	0	0	0.055	0	0	0	0	0	0	0
positive	positive	0	0	0	0.221	0	0.097	0.015	0.023	0	0	0
positive	positive	0.085	0.161	0	0.050	0.161	0	0.035	0.019	0	0.061	0.033
negative	negative	0	0	0	0	0	0	0.014	0.016	0.056	0.013	0.027
positive	positive	0	0	0	0	0	0	0.012	0	0	0.011	0.023
positive	positive	0	0	0	0	0	0	0.011	0	0	0	0
positive	positive	0	0	0	0	0	0	0	0	0	0	0
positive	positive	0	0	0	0.182	0	0	0	0	0	0	0

Fig.2. Sample of sentiment analysis results actual and predicted sentiment with Tf-idf and N-gram

From Kappa statistics, it is approved that naïve Bayes algorithm shows high degrees of acceptance for sentimental model as 0.71, 0.400, and 0.65 respectively as shown in table 1.

Table 1. Results of Sentiment Model with Naïve Bayes Classifier

Measurements	Yahoo Inc	Msft Inc	FB Inc
accuracy	**86.21%**	**72.73%**	**82.76%**
Kappa	0.717	0.400	0.65

The results proved that our proposed model achieved higher accuracy than the previous studies for the stock market news sentiment analysis. All the previous accuracies for stock market news sentiment analysis models do not exceed the range of 70 to 76%. In our proposed model. We obtained a range of accuracies ranging from 72.73% to 86.21%. Fig. 3 represents the correlation coefficient between the actual and predicted data labels as positive or negative.

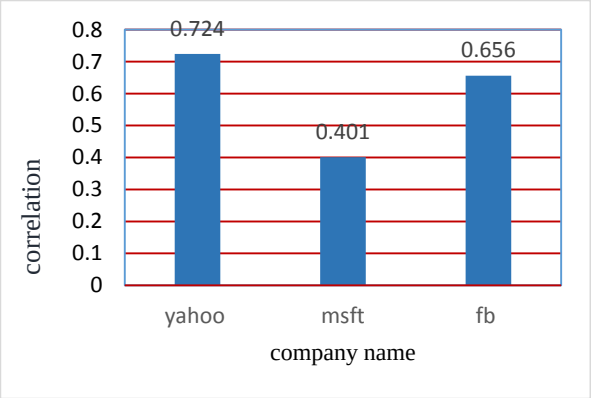

Fig.3. The correlation for the sentimental model Classification

Table 2. Results of performance of the model

	Prediction accuracy		
	K-NN	**SVM**	**Naive Bayes**
yahoo inc	75.86%	58.62%	86.21%
Msft inc	69.70%	66.67%	72.73%
Fb inc	72.41%	68.97%	82.76%

The correlation values demonstrated that the proposed model implies high degree of acceptance for news classification. The correlation values for three companies are 0.72, 0.40, and 0.65 for yahoo, Msft and Fb respectively.

For evaluating the sentiment analysis model based Naïve bayes classifier with Support Vector Machine (SVM) and K-NN algorithms. The results of the comparison are summarized in table 2. The table contains the results for the implementation of two classification algorithms with the prediction accuracy of our model and others, which are SVM and K-NN.

By comparing the prediction accuracy of the proposed model using naïve Bayes algorithm and other mentioned two algorithm, the comparison demonstrated that naïve byes algorithm outperforms the SVM and K-NN algorithms with textual data. Also, SVM has the lowest accuracy in dealing with textual data for our experimental dataset finding.

The performance of algorithm effects on the experimental dataset. Fig 4 represents the performance of other two classic data mining algorithms which are, K-NN and SVM on our excremental dataset. It was found that the different algorithms have an impact on the accuracy of the experimental dataset.

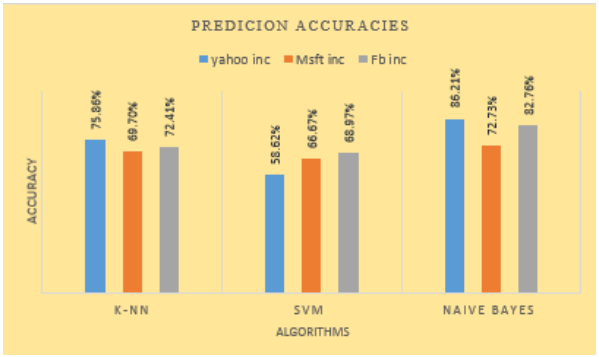

Fig.4. The comparison results of Sentimental Model algorithms with Naïve Bayes Classifier

B. Results of future stock trend prediction component

Table 3 shows the accuracy of prediction when applying our model using sentiment analysis and numeric data, compared to prediction accuracy when using sentiment analysis only. The obtained results are compatible with researches findings that, there is the strong relation between news and stock prices changes. All related studies depend on text analysis and use closing price in order to calculate trend or to compare trend results. The related studies achieved accuracies that do not exceed the range from 75% to 80% for the stock trend prediction.

Table 3. Results of Future Stock Trend Prediction Model Using K-NN Classifier

Measurements	sentiment attributes	sentiment &numeric attributes
accuracy	**59.18%**	**89.80%**
kappa	0.078	0.79

This study finds that depending only on the sentiment analysis for news in stock trend prediction produces accuracies starting from 59.18% up to 63%. Alternatively, when considering the numeric attributes represented in the open, high, low and closing prices, this enhance the prediction accuracy up to 89.80% for stock trend prediction. The prediction model depends on K-NN algorithm that performs well on textual and numeric data. The results demonstrate that our model is an effective way to improve the prediction accuracy of stock trend prediction with the higher accuracy. Kappa statistics demonstrates that using K-NN algorithm for stock trend prediction has a high degree of acceptance as 0.0789, 0.7879.

Our methodology is an effective way to improve the prediction accuracy of the stock market behavior based on sentiment analysis and historical numeric values. To verify this method, our proposed model is compared with the accuracies of previous studies results. Table 4 demonstrates that our proposed model outperforms the

other previous studies.

Table 4. Previous studies results for the stock market behavior perdition Compared with our proposed model accuracy

Previous Studies	Accuracy
L.L Bing et al. model [13]	76.12
Y.Cara et al. model [14]	60.39% :67.73%
Shynkevichl et al. model [15]	79.59%
Phyng model [11]	75%
y.Kim et al. model [19]	60% : 65%
our proposed model	**89.80%**

V. Conclusion

The proposed model investigated the simultaneous effect of analyzing different types of news along with historical numeric attributes for understanding stock market behavior. Our proposed model improved the prediction accuracy for the future trend of stock market, by considering different types of daily news with different values of numeric attributes during a day.

Three categories of news data were considered: news relevant to market, company news and financial reports that were published by financial experts about stocks. The proposed model consists of two stages, the first stage is to determine the news polarities to be either positive or negative using naïve Bayes algorithm, and the second stage incorporates the output of the first stage as input along with the processed historical numeric data attributes to predict the future stock trend using K-NN algorithm. The results of our proposed model achieved higher accuracy for sentiment analysis in determining the news polarities by using Naïve Bayes algorithm up to 86.21%. In the second stage of analysis, results proved the importance of considering different values of numeric attributes. This achieved the highest accuracy compared to other previous researches, our model for predicting the future behavior of stock market obtained accuracy up to 89.80%.

In the proposed model, both Naïve Bayes and K-NN methods lead to the best performance. The results of the proposed model are compatible with researches that state that there is a strong relation between stock news and changes in stock prices. This model can be updated in the future by including some technical analysis indictors, also we can consider the recognition of emotional sentences in determining news polarities, as well as the influence of news that appears in social media.

References

[1] B. O. Wyss, "Fundamentals of the stock market," p. 245, 2000.

[2] M. K. Jiawei Han, *Data MiningConcepts and Techniques*, Second Edi. Urbana-Champaign, 2006.

[3] K. Tan , steinbach, *Introduction to data mining*. 2006.

[4] W. Walter, K. Ho, W. R. Liu, and K. Tracy, "The relation between news events and stock price jump : an analysis based on neural network," *20th Int. Congr. Model. Simulation, Adelaide, Aust. 1–6 December 2013 www.mssanz.org.au/modsim2013*, no. December, pp. 1–6, 2013.

[5] A. Søgaard, "Sentiment analysis and opinion mining," ... *Lang. Comput. Group, Microsoft Res. Asia …*, no. May, 2013.

[6] K. J. C. Sahil Zubair, "Extracting News Sentiment and Establishing its Relationship with the S & P 500 Index," *48th Hawaii Int. Conf. Syst. Sci. Extr.*, 2015.

[7] C. Paper, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," *J. Emerg. Technol. Web Intell.*, no. October 2014, 2016.

[8] pasi tapanainen gregory grefenstette, "™ what is a word,what is a sentence?problem of tokanization," *maylan Fr.*, p. 9, 1994.

[9] H. D. Hana Alostad, "Directional Prediction of Stock Prices using Breaking News on Twitter," *IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, pp. 0–7, 2015.

[10] M. F. Patrick Uhr, Johannes Zenkert, "Sentiment Analysis in Financial Markets," *IEEE Int. Conf. Syst. Man, Cybern.*, pp. 912–917, 2014.

[11] P. M. Hoang Thanh, "Stock Market Trend Prediction Based on Text Mining of Corporate Web and Time Series Data," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 18, no. 1, 2014.

[12] S. M. Price, J. Shriwas, and S. Farzana, "Using Text Mining and Rule Based Technique for Prediction of," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 4, no. 1, 2014.

[13] L. I. Bing and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of A Company ' s Stock Price Movements," *IEEE 11th Int. Conf. E-bus. Eng. Public*, 2014.

[14] B. D. T. Yahya Eru Cakra, "Stock Price Prediction using Linear Regression based on Sentiment Analysis," *Int. Conf. Adv. Comput. Sci. Inf. Syst.*, pp. 147–154, 2015.

[15] Y. Shynkevichl, T. M. Mcginnityl, S. Colemanl, and A. Belatrechel, "Stock Price Prediction based on Stock-Specific and Sub-Industry-Specific News Articles," 2015.

[16] S. S. Umbarkar and P. S. S. Nandgaonkar, "Using Association Rule Mining : Stock Market Events Prediction from Financial News," vol. 4, no. 6, pp. 1958–1963, 2015.

[17] R. Desai, "Stock Market Prediction Using Data Mining 1," vol. 2, no. 2, pp. 2780–2784, 2014.

[18] I. Journal, O. F. Social, and H. Studies, "TIME SERIES ANALYSIS ON STOCK MARKET FOR TEXT MINING," vol. 6, no. 1, pp. 69–91, 2014.

[19] Y. Kim, S. R. Jeong, and I. Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 6, no. 1, pp. 1–13, 2014.

[20] S. S. Abdullah, M. S. Rahaman, and M. S. Rahman, "Analysis of stock market using text mining and natural language processing," *2013 Int. Conf. Informatics, Electron. Vis.*, pp. 1–6, 2013.

[21] U. States and E. Commission, "SECURITIES AND EXCHANGE COMMISSION THE 'TRANSITION REPORT PURSUANT TO SECTION 13 OR 15 (d) OF THE SECURITIES," vol. 302, 2014.

[22] N. L. More, C. T. Any, and O. U. S. Equities, "About NASDAQ."

[23] D. Lyon and B. Cedex, "N-grams based feature selection and text representation for Chinese Text Classification ZhihuaWEI," *Int. J. Comput. Intell. Syst.*, vol. 2, no. 4, pp. 365–374, 2009.

[24] salton and Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Inf. Process. Manag.*, vol. 24(5), p. 513–523., 1988.

[25] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining*. 2015.

[26] M. Mittermayer, "Forecasting Intraday Stock Price Trends with Text Mining Techniques," *Proc. 37th Hawaii Int. Conf. Syst. Sci. - 2004*, vol. 0, no. C, pp. 1–10, 2004.
[27] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," vol. 3, no. 5, pp. 605–610, 2013.
[28] Mr. B. Narendra and Mr. K. Uday Sai et al., "Sentiment Analysis on Movie Reviews : A Comparative Study of Machine Learning Algorithms and Open Source Technologies," IJISA, no. August, pp. 66–70, 2016..
[29] P. A. Idowu, C. Osakwe, A. A. Kayode, and E. R. Adagunodo, "Prediction of Stock Market in Nigeria Using Artificial Neural Network," IJISA, no. October, pp. 68–74, 2012.
[30] N. and K. J. Navale, "Prediction of Stock Market using Data Mining and Artificial Intelligence," Int. J. Comput. Appl., vol. 134, no. 12, pp. 9–11, 2016.

Authors' Profiles

Ayman E. Khedr is an Associate Professor, now he is the head of Information Systems Department, Faculty of Computers and Information Technology, Future University in Egypt.

Also, he is the supervisor of the Quality Assurance Department and a member of continuing education board. He has worked at the Faculty of Computers and Information, Helwan University in Egypt, and he has been the general manager of Helwan E-Learning Center. His research is concentrated around the themes (scientific) data and model management, data mining, Bioinformatics and cloud computing.

S.E.Salama is a lecturer, at Information Systems Department, Faculty of Computers and Information System, Helwan University in Egypt. Also, she is working a lecturer at Faculty of Computers and Information Technology,King Abduaziz University, Jeddah, Kingdom of Saudi. Her research is concentrated around the data mining and cloud computing.

Nagwa Yaseen Hegazy is a master student at Information Systems Department, Faculty of Computers and Information Systems, Helwan university.

How to cite this paper: Ayman E. Khedr, S.E.Salama, Nagwa Yaseen,"Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.7, pp.22-30, 2017. DOI: 10.5815/ijisa.2017.07.03

