# A Proposed Model for Predicting Stock Market Behavior Based on Detecting Fake News

Amira M. Idrees
*Faculty of Computers and Information, Fayoum University*

Mohamed H. Ibrahim
*Faculty of Computers and Information, Fayoum University*

Nagwa Y. Hegazy

*Faculty of Computers and Information, Helwan University*

ABSTRACT: Stock market is an important area of research due to its higher earnings. The higher earnings for the stock market also imply higher risks, so a large amount of data generated by the stock market is considered a treasure of knowledge for investors. There are several aspects that affect the stock market fluctuations the most important of them is news data. News data have an influential effect on the investors' thoughts and beliefs. Using machine learning and textual data processing considered a significant part of the stock market analysis. Researchers concerned with designing the suitable model to predict the future behavior of stock market to avoid investment risks. It was found that there is a strong relationship between stock news and changes in the stock prices. This study aims at proposed a framework for detecting the stock market fake news that helps in avoiding higher investment risks and improve the stock market prediction accuracy. Discovering the best combination of machine learning algorithms that lead to the best performance of the prediction model that designed based on news sentiment analysis and numeric data analysis. Different experiments have been applied to uncover algorithms that led to the best performance and raising the prediction accuracy up to 92%.

**Keywords:** Stock Market, Fake News, Sentiment Analysis, Text Mining, Machine Learning, Random Forest.

## 1 INTRODUCTION

The stock market has many factors effects on the market behavior; news data is the main factor of them. Detecting news credibility and authenticity considered an important area of research and has lower interest in the stock market researches. Incorrect news content led to more fluctuation in the stock market through many buying and selling signals that cause the large financial losses misleading content of the stock market news affects the investor's behavior and decisions. Social media and knowledge sharing platforms are large-scaled as a source of information for investors this information includes the stock market news data as well as the tremendous number of social media users that biases their thoughts and decision toward specific stocks. In the stock market the coast of information is very high so, detecting news credibility is a major task to avoid the large financial loses. Investors usually wish to achieve higher profits on their investments through determining which stocks and the best time to buy or sell; this is achieved by designing an accurate prediction model for the stock market behavior. Machine learning techniques used to explore the stock market pattern. Machine learning includes supervised and unsupervised approaches (Witten,Frank,Hall2011), (Kaseb, Khafagy, Ali, and ElSayed. 2018). Text mining is a process of handling the unstructured data and considers a step of knowledge discovery. Text preprocessing techniques includes tokenization, stemming and stop word removal (Vijayarani and Janani 2016). Sentiment analysis is the process of determining people's attitudes, opinions, evaluations, appraisals, and emotions towards entities such as products, services, organizations, attributes using NLP, statistics, or machine learning methods from text data (Patrick,Zenkert 2014).

## 2 RELATED WORK

There is a number of researches that worked in the area of prediction in general (Sahal, Khafagy, and Omara , 2016), and in the stock market prediction in specific, some of them ap-

proached to predict the future price based on historical stock prices such as (Witten, E. Frank, and M. a. Hall, 2011), (Sahaj& Govinda 2017), (L. I. Bing and C. Ou, 2014 ) others approached to predict the stock market behavior based on analyzing news sentiment such as (SVijayarani, Ilamathi, and Ilamathi, 2016), (Tan , steinbach, Kumar, 2006), (M. M. S. and P. A. K. Senthamarai Kannan, P. Sailapathi Sekar, 2010), and other studies aimed at building their prediction models based on news and some of historical stock prices such as ( S. M. Price, J. Shriwas, and S. Farzana, 2014). The works in literature is represented that studies for the stock market prediction such as (Sahaj& Govinda 2017), (Mazen et al. 2018),(Khedr,Salama,and N.Yaseen 2017) others targeted fake news detection such as (Graink and mesyura 2017). In the field of the stock market for fake news detection, there are few studies till now works to detect the credibility of stock market news in spite of its importance and impact for investment decision, investor 's behavior. Kogan et al. (2017) and Martin et al. (2017) proposed a model for fake news detection in financial markets.

## 3 METHOD

Fake news detection considered an important task for stock market prediction because it's effect on the investor's thoughts and believes. It also has a great effect on the stock market prices according to a lot of works of literature that proved that there is a strong correlation between news releases and stock market prices. The aim of our proposed model is detecting the stock market fake news and then enhancing the prediction accuracy of the stock market through combing stock market news and historical stock prices. The analysis of stock market news is based on analyzing different types of news for every day and determines their effect on the stock market along with historical stock prices (OHLC). Our proposed model implemented based on three steps: The first step, considers detecting fake news and then filtering it from our dataset to avoid nonfactual news thus avoid unauthenticated sources of information that cause stock market fluctuation and higher investment risks to improve the performance of the stock market prediction model. The second step, comparing different machine learning algorithms to

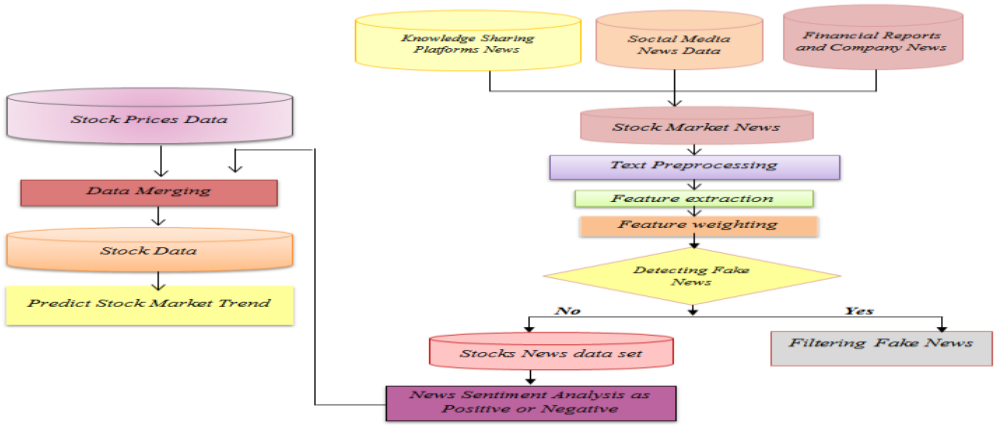


Figure-1 proposed model

discover the best combination of algorithms that lead to the enhanced performance of the stock market news sentiment analysis as positive or negative news. Third step, discovering the best combination of machine learning algorithms to enhance the performance of the prediction model that designed based on news sentiment analysis and numeric data analysis to predict the stock market future behavior as fall or raise.

### 3.1 *Data description:*

Our proposed model implemented based on united states market data. NASDAQ is the largest stock market in United States that combines stocks for largest companies in United States. The stocks for yahoo.inc, Facebook Inc(Fb), Microsoft corporation(MSFT) have been used in our experiment. We were collected different types of news along with historical stock prices for each company. For news dataset, we collect news from different data sources which of them

from authenticated sources such as wall street journal, Reuters, companies websites, Google finance, yahoo finance, Nasdaq.com, ecomomics.com and share market update. The other part of news dataset is collected from social media news (Twitter) along with stock mews from knowledge sharing platforms such as seeking alpha and motley fool. The second type of dataset is the historical stock prices for the focused companies stocks that include opening and closing prices along with the highest and lowest stock prices for the company stocks during a day.

### 3.2 *Description of the proposed model phases*

#### *3.2.1Fake news detection model:*

Fake news considered an important task to determine factually incorrect and misleading news for investors. Stock market fake news target to affect the investor's opinions and decisions about their investment portfolio so, it can cause large financial loses. Detection fake stock market news model is shown in figure 1. For the collected stock market news from different data sources some of them collected from authenticated sources and others collected from different knowledge platforms such as seeking alpha and motley fool. The performed Text preprocessing techniques includes using tokenization, stop words and stemming and text normalization. Text normalization techniques have been applied for news corpus by transforming different forms of text into common standard format through transforming all letters in news into lowercase. After that N-gram has been used for news corpus as a syntactic analysis technique in order to extract features from news corpus as a serious of tokens for length N. Our model generated bigrams features from news corpus. N-gram has a robust performance in extracting features from text because of automatic capturing for the most frequent roots in news data And good representation that is provided by n-gram, does not require using a specific dictionary, As well as its tolerance for spelling errors (D. Lyon and B. Cedex, 2009).

Vector space model has been used for document representation to represent terms as a vector model and capture the extracted features form stock market news corpus. TF-IDF(term frequency-inverse document frequency) is a feature weighting method used for weight each term in news articles to indicate the word importance in news documents or in the collection of corpus. TF-IDF used to replace each token with weighted value (K. S. Loke, 2017). TF is the term frequency for term t in document d, that weights term based on the occurrence of a term in the document. IDF weight is based on the documents that contain the term t in the collection of news articles.

After features are extracted and weighting Random forest classifier has been applied to classify the stock market news as fake or correct news. Our proposed model has two main classes "fake" and "correct" news. Fake news detection model is implemented using Random forest classifier. Random forest one of supervised classification machine learning algorithms which is an ensemble classification method. This technique builds multiple decision trees and then, combines these decision trees to produce random forest (A. Assaf and E. Alnagi, 2013), (M. N. Elagamy, C. Stanier, and B. Sharp, 2018). The output of several independent decision trees is combined and the majority vote used to produce the optimal predictive model. For the stock news data, random forest builds multiple decision trees and combines the output for these trees based on the majority voting to produce the optimal classifier and determine to obtain class label (which news articles are fake and correct). Random forest has been implemented for detecting fake or correct stock news. Random forest classifier is extremely flexible and has higher accuracy. We noticed that for fake news data there are common features to discriminate fake news. It is usually have shorter news content lower authenticity scores.

#### *3.2.2News Sentiment Analysis Phase:*

In this phase we perform the following steps for text normalization techniques such as along with previous text prepressing techniques that performed for news data in the fake news detection component. Sentiment analysis for stock market news to be either positive or negative for the stock market behavior has been implemented using different machine learning algorithms and compare their performance to uncover which of them have higher performance and more suitable for the complex nature of stock market news dataset. We explored various machine

learning methods such as support vector machine (SVM), K-nearest neighbor (k-NN), naïve Bayes and logistic regression in order to improve the classification accuracy. SVM is one of the classification methods that has its roots in statistical learning theory. SVM is based on the idea of separating two datasets by enforcing a margin and train to find the maximum margin. It also, find the distance between the point and hyperplane that has the maximum distance of the closest point to the margin called support vector. Naïve Bayes is based on conditional Independency between features.

Logistic regression one of the popular classification algorithms that are used logistic function to estimate the probabilities between data labels (as positive or negative) and the extracted features form news data. This means that the logistic function represents that the probability of class label occurrence is a linear function combination between independent predictor variables and the extracted features. The classifier used logistic function to calculate the likelihood that "positive" class label will occur with the specific group of features and will not occurs with another group of features that belongs the "negative" class label. Logistic regression considers the dependency between features and performed well with binary classification.

### *3.2.3Historical Stock Prices Analysis phase:*

This phase includes analyzing the stock market historical stock prices that consider the daily opening, closing, high and low prices for three companies; Yahoo Inc, Facebook Inc, and Microsoft Corporation respectively. For each company, we have a historical opening, closing, high, and low continuous values. Preprocessing technique have been performed for historical stock prices by converting all continuous values into discrete values as in (Khedr,Salama,and N.Yaseen 2017). The output data of this phase will be in the prediction model after merging it with news sentiment analysis data.

### *3.2.4Predicting the Stock Market Future Behavior Phase*

In this phase, we merging the output of news sentiment analysis after detecting fake news and filtering it form news dataset in this phase we merging data news polarities data as positive or negative news with the output of the third phase historical stock prices analysis. Different types of news have been considered for news polarities. More than two daily news articles are considered. For the prediction model, we compare the performance of three machine learning classification algorithms in order to enhance the performance of the prediction model for the stock market future behavior as 'falling' or 'raising' signals and to find the best combination of machine learning algorithms for news sentiment and stock market prediction that led to enhanced performance. SVM, Random forest and k-star algorithms have been used to predict the future behavior of the stock market. K-star algorithm is a type of lazy learners that delays the training set until there is a need to test data instances.

## 4 RESULT AND DISCUSSION

In this section, we will discuss the experimental results that represented in three sections; the first section represents results for fake news detection through the sentiment analysis for news content as fake or correct news, second section concerned with the experimental results for stock market news sentiment as positive or negative effect on the stock market or on the investors decisions with different machine learning algorithms to find the best performance, finally the third section describes the results for predicting the stock market future behavior as falling or raising signal.

1. *Results for detecting the stock market fake news:*

The experimental result demonstrated that it has a strong effect on the stock market behavior because of their impact on investor's believes and decisions. Random forest algorithm demonstrated a reasonable accuracy for fake news detection. Random forest algorithm used to detect the credibility of the stock market news as fake or correct news for three companies yahoo inc, FaceBook Inc and MICROSOFT. The results for news credibility founded that achieved accu-

racies are 66.6%, 69.2%, 61.1% for Yahoo Inc, Facebook inc and MICROSOFT respectively. The model performance measured by kappa statistics and the results demonstrated that the proposed model is acceptable. It was found that from our experimental results fake news has general features such as it usually has shorter content, fewer self-references, higher insight words such as think, know and discrepancy words such as (think, should) this features are compatible with previous studies for fake news detection models.

2. *Results for the stock market news sentiment:*

In this phase the stock market news sentiment analysis performed based on different algorithms to find the best performance. The overall experimental results are represented in figure 2. Empirically it has been shown that logistic regression achieved higher accuracy than K-NN, Naïve Bayes and SVM algorithms; this means that logistic regression more suitable for textual data analysis than others. From figure 2 logistic regression archived accuracy for the sentimental model up to 88.24%, 86.21% and 81.82% for Yahoo, FB, and MICROSOFT respectively. The obtained higher accuracy for logistic regression because of it considers the dependency between features and has higher performance in the case of binary classification. The experiment includes using the random forest for news polarities but it does not achieve reasonable accuracy because of the large size of data than the size of data for fake news detection and the sentimental features that does not enough to give a label. The feature evaluated as split decisions are not informative and therefore produce imbalanced decisions as well as the complexity of building multiple classifiers on huge text data.

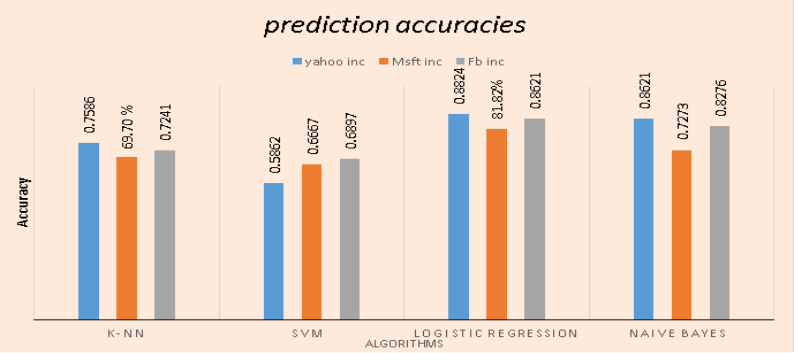


Figure 2 Results for Algorithms Performance for News Sentiment Analysis phase.

*3. Results for predicting the stock market behavior phase:*

In this section, we describe the experimental results for the stock market prediction model for the future behavior as falling or raisin signals. Performance of different algorithms has been compared. The results for all are promising. The performance of three supervised learning have been comparing the algorithms are; SVM, Random forest and k-star algorithm. It has been found that; Random forest algorithm has higher performance than others it achieved accuracy up to 92.3% for MICROSOFT Company. The training and testing files have been shuffled to measure the validation of our proposed model.

The experimental results demonstrated that the model based on random forest outperformed SVM and k-Star with accuracy up to 83.9%, 92.3%, 91.3% for yahoo, MICROSOFT inc, Facebook inc respectively. The achieved accuracies for predicting the stock market future behavior model based on SVM algorithm achieve accuracy up to 82.9%, 84.6% and 79.2% for yahoo, MICROSOFT inc, Facebook inc respectively, finally the prediction model based on k-star algorithm performance also was applicable with accuracy up to70%, 80.6%,82.9%.

The experimental results proved that our proposed model is an effective way to predict the stock market based on stock market news detection, analyzing different types of news and historical stock prices. Our model has been tested using kappa statistics to compare the observed accuracy with the expected accuracy to determine the degree of acceptance and approval for the proposed model. The observed values from figure 2 demonstrated that; our mythology achieved a higher degree of acceptance using Random forest algorithm.

## 5 CONCLUSION

Stock market prediction considered the most essential area of research because of the higher earnings and its importance for countries economic growth. There are several factors affects the stock market. News releases considered a significant factor that causes the stock market fluctuations because of its great impact on inverse's thoughts and decisions toward their investments. Detecting the news credibility is a new growing area of research In the stock market prediction so, detecting the authenticity of news have a great impact in improving the stock market prediction and avoiding the large financial losses. The proposed model detecting the stock market news releases using random forest algorithm with reasonable accuracy. News sentiment analysis as positive or negative has been performed using logistic regression. Logistic regression is more suitable for sentiment analysis data through the comparison between four popular machine learning algorithms by achieving accuracy up to 88.24%. Comparison between three machine learning algorithms has been performed to find the best combination of algorithms that led to higher prediction accuracy for the stock market. Random forest algorithm achieved the best performance and higher accuracy than other algorithms. For future work, the model can be enhanced by consider larger size of fake news and consider some technical analysis features.

## 6 REFERENCES


Abu Assaf, E. Alnagi, Al-radaideh. 2013 Predicting stock prices using data mining techniques 1, Int. Arab Conf. Inf. Technol., pp. 1–8,.

B.D.Trisedya, Y. E. Cakra. 2015 ―Stock Price Prediction using Linear Regression based on Sentiment Analysis,‖ Int. Conf. Adv. Comput. Sci. Inf. Syst., pp. 147–154.

C. Science. 2016, Text Mining : Open Source Tokenization Tools – an Analysis, Adv. Comput. Intell. An Int. J., vol. 3, no. 1, pp. 37–47.

D. Lyon and B. Cedex. 2009 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.

G. Sahaj. 2017 Stock Market Prediction Using Data Mining 1, in 2017 International Conference on Intelligent Sustainable Systems (ICISS), vol. 2, no. 2, pp. 2780–2784.

Khedr AE, S.E.Salama, Yaseen N. 2017 Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. Int J Intell Syst Appl.;9(7):22-30. doi:10.5815/ijisa.2017.

K. Insights. , 2017 Fake news, social media and the value of credible content, no. 2, pp. 1–11.

K. Tan , Steinbach. 2006 Introduction to data mining.

K. S. Loke. 2017 Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests, pp. 38–42.

Kaseb, Mostafa & Khafagy, Mohamed & Ali, Ihab & M. Saad, ElSayed. 2018 Redundant Independent Files (RIF): A Technique for Reducing Storage and Resources in Big Data Replication, 182-193. 10.1007/978-3-319-77703-0_18.

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

M. M. S. and P. A. K. Senthamarai Kannan, P. Sailapathi Sekar. 2010, Financial Stock Market Forecast using Data Mining Techniques, Int. multiconference Eng. Comput. Sci., vol. I.

M. Granik and V. Mesyura. 2017, Fake news detection using naive Bayes classifier, Electr. Comput. Eng. (UKRCON), 2017 IEEE First Ukr. Conf., pp. 900–903.

M. N. Elagamy, C. Stanier, and B. Sharp. 2018, Stock market random forest-text mining system mining critical indicators of stock market movements, 2nd Int. Conf. Nat. Lang. Speech Process. ICNLSP 2018.

R. Desai. 2017, Stock Market Prediction Using Data Mining 1, in 2017 International Conference on Intelligent Sustainable Systems (ICISS), vol. 2, no. 2, pp. 2780–2784.

Radhya Sahal, Mohamed H. Khafagy, and Fatma A. Omara. 2016, Comparative Study of Multi-query Optimization Techniques Using Shared Predicate-based for Big Data, International Journal of Grid and Distributed Computing Vol. 9, No. 5.

S. M. Price, J. Shriwas, and S. Farzana. 2014, Using Text Mining and Rule Based Technique for Prediction of stock market price,Int. J. Emerg. Technol. Adv. Eng., vol. 4, no. 1.

Salton and Buckley. 1988, Term Weighting Approaches in Automatic Text, Retrieval,‖ Inf. Process. Manag., vol. 24(5), p. 513–523.

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

Witten, E. Frank, and M. a. Hall. 2011, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, vol. 54, no. 2