Turkish Journal of Computer and Mathematics Education Vol.12 No. 5 (2021), 1022-1030

Research Article

Stock Market Forecasting Model From Multi News Data Source Using a Two-Level Learning Algorithm

El Bousty Hicham[a], Krit Salah-ddine[b]

[a]Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Laboratory of Engineering Sciences and Energy, Agadir, Morocco. E-mail: hicham.elbousty@edu.uiz.ac.ma
[b]Professor, Computer Sciences and Physics, Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir, Morocco. E-mail: s.krit@uiz.ac.ma

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Stock prediction retains the attention of a large part of the community. The emergence of new indicators mostly extracted from the web makes this domain of research challenging and in a continuous evolution. The present work tries to address the question of how to model financial news from multi-data source for the purpose of forecasting stock movement. We combined different news sources to enhance the accuracy of stock movement prediction. Data are collected from four financial news websites and proceeded individually by Support Vector Machine (SVM) algorithm then we aggregate outputs using an Artificial Neural Network (ANN) algorithm. Experiments were conducted and the results have shown that the designed two-level learning SVC&ANN algorithm has achieved better accuracy than simple news analysis models using a single information source.

Keywords: Machine Learning, SVC, Neural Network, Stock Forecasting, Bag of Words.

1. Introduction

Since the correct prediction of the market movement is rewarding for investors and traders, they are in permanent search for new models, systems and indicators. It started from simple historical stock value analysis to volume analysis to financial result reports to technical indicators to events analysis and finally the social network analysis. However, in efficient market the stock evolution is unpredictable. Fama distinguishes three forms of market efficiency [1]:

- Weak efficiency: available information contains only historical price and volume.
- Semi-strong efficiency: contains all public information included annual reports, dividends announcements.
- And strong efficiency: besides the public information it is enriched with private information.

For him None of the forms above allow abnormal benefits and no one can beat the market. He claims that the price already includes all available information. Hence operators have the same degree of information. Since Fama supposes that operators are rational, accordingly they will make the same movements and will try to buy and sell the same shares. This result the dissolution of the market and no exchange will be possible.

Some scientists refute this supposition and consider two categories of market agents: informed agents and non-informed agents. The firsts acquire information with certain cost and they act according this knowledge, whereas uninformed agents decrypt data from the price. The latter doesn't reflect all available information, otherwise informed agents will no more pay for information acquisition. Others demonstrate that there is a fundamental price that the market price trends generally towards this value. This phenomenon is known as mean reversion. Poterba explains that if the market price and the fundamental value diverge the speculative forces eliminate these differences [2]. This make the market predictable and give opportunity for stockholders.

Some Academics put the different forms of market efficiency under test, such as Poterba and Summers who show autocorrelations in the weak form (historical data) especially in the case of short-term predictions [3]. For dividend information publication, Charest shows that significant residuals were observed in the month following dividend changes [4]. He concludes that the market was slow in digesting dividend information. Symmetrically, el bousty & al. showed that the best prediction accuracy was four days after information publication [5].

The efficiency of the market depends roughly on the information flow and the human behavior. Any failure in the information process induce market efficiency anomaly. Behavioral finance theory, claims that decision-making depends on human psychology. The individual cognitive biases psychological and heuristic variables

determine his comportment and strategy [6]. These parameters may involve an irrational behavior and the market strengths can't compensate the human failure. Hence human irrationality maybe another source of market inefficiency.

All the work has been done up to now disproving the market efficiency give reason to the development of some trading strategy, in particular fundamental and technical approaches. The first one is based on the analysis of available information about the company strength, the markets and the economy in general. It is interested in evaluating its real value according to a set of macro and micro economic parameters. On the other hand, the technical analysis is mainly compiling the evolution of historical prices and trading volumes to estimate values and trends. It is largely based on the graph's analysis.

Machine learning algorithms is used to enhance the accuracy of these strategies. Powerful algorithms are combined with technical and/or fundamental indicators and huge amount of data are analyzed thanks to mutualisation of cloud resources. First uses of machine learning in financial prediction deals with structured data such as prices, volume, financial reports…. Now all information, which is often unstructured, are combined to empower the forecasting. This raised more complex challenges especially how to process text and even non-written information to uncover hidden knowledge.

This work is an attempt to build a model for forecasting shares from multi news data sources and using two-level machine learning model. The motivation behind this work is twofold:

First, Comparing the accuracy of a single news data source model with a multi news data source model and second, inspect the behaviour of our designed model when varying the number of data sources.

The rest of this work is organized is as follows. Section 2 introduces some previous research work on predicting stocks through text analysis. Section 3 presents method used for retrieving and preparing datasets. Section 4 describes the proposed model. Section 5 depicts realized experiments. Sections 6 shows the results. And finally, Section 7 concludes the contribution of this research work.

2. Related Work

Stock prediction through machine learning algorithms is an active research area. Some research exploits the historic price trading data (open, high, low & close prices) in stock predictions. Jigar Patel.et all computed ten technical parameters and compared the performance of four machine learning classifiers [7]. Results show that random forest outperforms Artificial Neural Network (ANN), Support Vector Machine (SVM) and naive-Bayes algorithms. David M. Q.Nelsonet all used Long short-term memory (LSTM) algorithm to predict future trends of stock prices from historical prices and some technical analysis indicators [8]. Volume can also be useful in stock movement forecasting as shown by Edson Kambeu [9]. The trading volume for the third previous day influence current stock market index movement at the Botswana Stock Exchange. All the above technics exploit structured data only, but as the web information grows continuously, there is an increasing need for handling unstructured datatoo. Studies showed that extracted data from web has great influence on stocks movements direction [10]. Indeed, news events and social media data were inspected on multiple studies to unlock insights hidden within texts [5, 11].

The observed influence of news events and social media data on stock movements along with technical prediction methods led to advanced research combining two or more of these technics. Nisal Wadug and Upkesha Ganeguda [12], suggested four components model for better results in stock prediction. Keyword Extraction Module that extract macroeconomic indicators from published financial reports using crawlers or OCR technology. The second component is the Incident Mining Module which focuses on gathering data from newspapers and social media. News relevance is measured through Impact Analysis Module, Google trend can be used to determine the importance of extracted events. The last component is the Performance Isolation Module, that attempts to separate the exact performance of a company from external effects. Line of studies has demonstrated that the multi-source information outperforms each prediction based on a single source [13], including the work conducted by Xiadong Li et al. that compared a news-based model, historical

Table 1. Example of Journal Data Frame

Date	Title	Corpus	Share
04 February 2019	Financements verts : une ligne de 20 millions d'euros pour la bmci auprès de la berd	La bmci a signé le 4 février 2019, un contrat de partenariat avec la banque européenne pour la reconstruction et le développement (berd) pour le programme geffmorocco, en sa qualité de "leader dans ce segment", indique la banque dans un commu...	bmci
04 February 2019	Attijariwafa bank: africaine de bourse deviant "attijari securities west africa"	Africaine de bourse, société de gestion et d'intermédiation du groupe attijariwafabank dans la zone uemoa (union économiqu….	attijariwafa
01 February 2019	Mario camacho poursuit ses achats sur cartier saada	Cartier saada veut prospérer sur le marché américain mariocamachoinc, un des leaders de la distribution des olives de tables aux etats unis d'amérique, spécialisé dans les ve....	cartier saada
01 February 2019	Alliances : dernière ligne droite pour le reprofilage de la dette privée	Alliances vient de convoquer les porteurs de ses obligations en assemblée générale 26 février afin d'approuver le principe du remboursement des obligations et du paiem…	alliances

Algorithm 1 Data source construction

```
Input
    J    Web journals
    C    Companies listed in Casablanca Stock Exchange
Output
    DS_j Data source for journal j
for j in J do
    for article in j.articles do
        for c in C do
            if c in article then
                add article, c to DS_j
            end if
        end for
    end for
end for
```

Figure 1. Data Source Construction Algorithm

price-based, naïve combination of these two sources and a Multi kernel learning model combination [14]. MKL performs better on most tests and accuracy of news-based model is similar to naïve combination model. Aparna Nayak combined all available data about selected companies especially historical data, news and tweets [15]. First, continuous trend pattern is calculated from the last three days prices (1 if the last three days trend continue in the same direction else it's equal to 0) then volume variation is compared to the trend at the same day and volume variation pattern is established. These two patterns are combined with polarity extracted from news and tweets. The daily predictions achieved about 70% accuracy using Boosted Decision Tree Model.

Although news events and tweets proof its influence on stock price movement they are rarely inspected alone without combination with historical data or technical indicators. One of the few researches examining impact of published events on stock evolution is the work realized by Aditi Kaushal and Prerit Chaudhry [16]. They scraped news articles that are linked to the AAPL share from Reuters website, and assigned them a value of 1 or -1 depending on the sentiment reflected by the article (positive/negative). They get a general sentiment for a specific day by summing all sentiment values for the same day. Authors believe that not all news has the same impact on the stock, they consider then another parameter which is the magnitude of the news, it's calculated by multiplying the Google trend value of 'Apple' in specific period by the general sentiment calculated previously. The forecasting of the current day movement involves the magnitude values of the last 14 days. The influence of previous days is decreased exponentially using a decay function. These values are fed into SVM, Logical Regression and Naïve Bayes algorithms and comparison showed that the best accuracy results are obtained using SVM algorithm.

Most researches focus on analyzing news articles scraped from one single data source or combining news analysis with other indicators (historical data, tweets, financial reports, technical indicators), no one tries to

aggregate multiple news data sources to predict accurately stock movements. This article suggests a different approach for predicting stock movement by aggregating news collected from four different data sources using Artificial Neural Network.

3. Data Retrieval

In this work we considered articles collected from four different Moroccan economic journals between 15 September 2014 and 20 February 2019. The process of extraction is depicted in Figure 1. We developed a python utility for extracting articles' links from Journals' websites and then fetches news articles available on those links. Each article is inspected for the purpose of retrieving the companies is talking about and only articles that are linked to one of the listed shares in Casablanca Stock Exchange is kept. In parallel, we collected historical price for those stocks from Investing.com website. At the end of this process, we construct data frame for each economic journal that contains, corpus of collected articles, the title of the article, date of publication, concerned stock, trend at the specific day (1 for up and -1 for down). The shape of one of the journals data frame is presented in the Table 1. The final Input data frame is constructed by aggregating the previous four data frames. Indeed, each row contains the date, the concerning share, the trend column and articles' titles and corpuses published that day for this share Table.2. We observed that for some events, only one journal has published an article about it, whereas for some other events two or more articles were published. Publication of the news in only one journal could be considered as a signal of non-relevance of that event. Hence, all events that were published in one journal were eliminated.

4. Proposed System

4.1. Data Pre-Processing

Before that we can get insights from extracted articles, we should first preprocess these articles. Preprocess phase consists of tokenizing, removing stop words and vectorizing. During the data processing phase each data source is treated separately (each journal is represented by a column in the aggregated input data source) Figure 2.

4.1.1. Text Tokenization

In order to apply any natural language processing technic, text is generally split into tokens (words in our case), then POS tag, named entities or any other technic can be easily applied. The main goal of this operation is eliminating punctuation and storing separately each word.

4.1.2. Stop Word Removal

The pre-processing starts with removing stop words from the corpus. Numbers, words composed from few characters (less than three characters) and those present on most articles are all removed from tokens list. This step is most useful when applying a Bag of words technic in order to reduce the features space.

4.1.3. Vectorization

Vectorization is the process of calculating the occurrence of words in the feature space. Each article corresponds to an occurrence vector and the matrix is made by regrouping all these vectors. The input Matrix X for a data source looks like below:

- $X=\begin{bmatrix} 0 & 2 & 0 & \cdots & 3 & 0 & 1 \\ 0 & 0 & 1 & \cdots & 0 & 1 & 0 \\ 1 & 0 & 0 & \cdots & 2 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 4 & 1 & \cdots & 1 & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 2 & 0 \end{bmatrix}$

X has as many columns as words in the feature space for that data source.

4.2. Designed Model

Table 2. Aggregated Data Source

Date	Journal 1 Article	Title	Journal 2 Article	Title	Journal 3 Article	Title	Journal 4 Article	Title	Share	Trend
10/02/2016	Le lancement du club afrique développement éma...	Afrique : entretien avec jamal ahizoune, dga d...	La4ème édition du forum international afriq...	Forum afrique développement : plus de 1.200 op...					attijari wafa	1
03/12/2015	Le conseil déontologique des valeurs mobilière...	Emission d'un emprunt obligataire subordonné d...	Les fonds collectés par le biais de l'émission...	A quoi va servir léemprunt subordonné déattija...	Attijari wafa bank envisage l'émission de 10.00...	Attijari wafa bank entend lever 1 milliard de dh			attijari wafa	-1
25/05/2015	Casa view du groupe klk immobilier, fal el hna...	Neuf projets immobiliers décrochent le label fnpi	Le groupe alliances annonce dans un communiqué...	Alliances mène des discussions avec ses créanc...			Alliances : le plan de restructuration est bie...	Le nouveau business plan 2015-2017 ne se...	alliances	1
11/05/2015	La société est correctement valorisée compte t...	Les titres total maroc offrent des perspective...	Dans ce septième numéro de l'hebdo des marchés...	Hebdo des marchés : total maroc arrive en bourse					total maroc	1

Algorithm 2 Matrix X preparation

Input
DS_j Data source for the journal j

Output
X_j Data source matrix for j

for *ds* in DS_j **do**
  Tokens=Tokenizing(ds)
  $FeautureSpace_j$=StopWordsRemoval(Tokens)
  for *article* in *ds.articles* **do**
    L=Vectorizing (article, $FeautureSpace_j$)
    append row L to X_j
  end for
end for

Figure 2. Data source Matrix Preparation

The analysis of a published article is not such as to accurately predict the evolution of stocks, event relevance is also decisive. The impact of events is usually measured through Google Trend [13,16]. However, it only allows us to evaluate the search frequency for a word (brand in the context of our research). Indeed, Google Trend evaluate the potential of price change but not the direction of the movement. In other words, any mistake in assessing the published article implies a wrong prediction. This is precisely what prompted us to develop an original model for quantifying the importance of published events. News impact is no longer estimated through Google Trend but it's evaluated by the number of data sources publishing this event.

As presented previously we considered four different journals for testing this model and this can be extended to unlimited data sources.

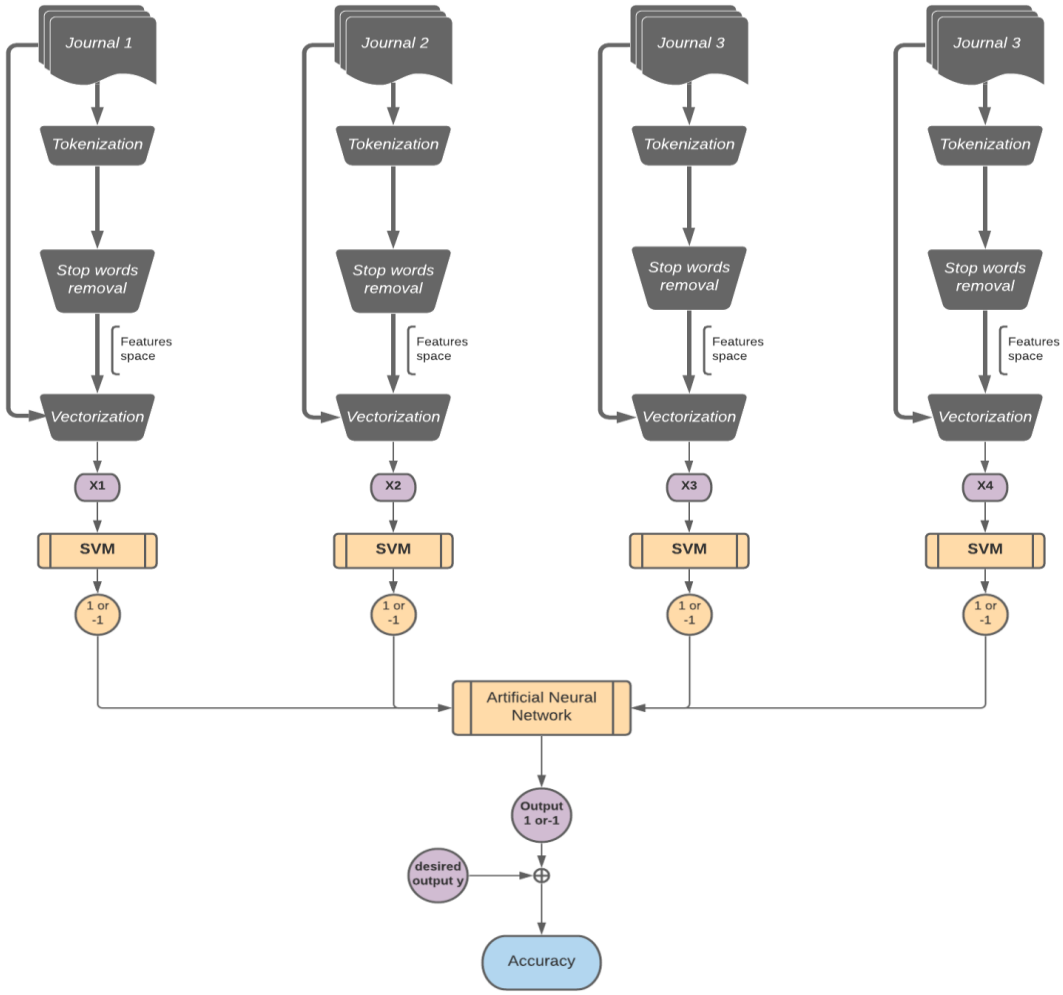

Figure 3. Designed Model

From the data processing step, we generate four different Matrixes X_j, one for each data source (articles columns in the aggregated data source). The trend column in the aggregated input data corresponds to the desired output y that has the shape below.

$$Y= = \begin{bmatrix} -1 \\ 1 \\ 1 \\ \vdots \\ -1 \\ 1 \\ -1 \end{bmatrix}$$

Our model is built from two levels (Figure 4). The first inspect the correlation between Input Matrix $\mathbf{X_j}$ and the trend **y** for each journal. SVM algorithm is widely used for text analysis, hence we opted for this algorithm to predict stock movement from the published article. This process is applied to each journal for the purpose of forecasting separately the stock movement. This technic is a kind of confirmation of the price change direction. A stock is more likely to go up when most journals estimated a positive change in the price. The four binary (1 or -1) values got from the first level are fed into an artificial neural network to correct SVM predictions Figure3.

Algorithm 3 Tow level predicition

Input

X_j Data source matrix for j

y Desired trend

Output

y_p Predicted trend

for j in 1..4 **do**

$pred_j$=SVM(X_j)

end for

y_p=ANN($pred_1$,$pred_2$,$pred_3$,$pred_4$)

Figure 4. Designed Model Algorithm

depicts the steps of the designed model which is deployed on python environment as a customed classifier.

5. Experiments

The aim of our work is first to inspect whether the use of multi-web news sources enhance the accuracy of stocks predications and second compare the reaction of our designed algorithm to different number of web news sources (two, three and four sources). Hence, we compared the accuracy of single data source with four data source model and then we checked out the evolution of the accuracy when varying the number of data sources. A part of the data is used for tuning the algorithms. Indeed, different values of C, Gamma, kernel, alpha and the max iterations parameters are tried and the best parameters are retained for the rest of the experiments.

For the algorithm training and testing the cross-validation approach was applied. Data were split to five folds and each time four are used for training and the fifth is fed into the algorithm for testing. Besides for the second experiment we also tried the 80%-20% split (80% for training and 20% for test) and 70%-30% split.

6. Experimental Results

First results showed that the accuracy of stock predictions from one single data source is around 60% (Table 3). This is in line with results obtained in [5]. The aggregation of the four data sources enhanced considerably the accuracy. It attends 86 % for the 80/20 experiment. The use of four different web journals for prediction helps in measuring the relevance of a published article. Hence news published about a brand in four journals is more likely to influence the stock movement than published news in a single journal. Besides, in text and sentiment analysis the way an article is written (style, words, event or opinion article…) is decisive in forecasting. Indeed, this technic aims to confront the predictions from the four web sources and the more likely prediction is selected through a trained ANN. From the second experiment we observed that the accuracy is increasing with the incorporation of new data source. It rose from 81% for two sources to 86% for the four journals (Table 5).

Table 3. One Journal Model Accuracy

Journal	Parameters			5-fold Accuracy					Mean Accuracy
	C	Gamma	Kernel	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Journal1	1	0,01	rbf	0,62	0,59	0,65	0,55	0,57	0,60
Journal2	1	0,01	rbf	0,55	0,55	0,59	0,57	0,53	0,56
Journal3	1	0,1	rbf	0,70	0,60	0,63	0,64	0,65	0,64
Journal4	10	0,001	rbf	0,53	0,61	0,52	0,51	0,64	0,56

Table 4. The Two-Level Algorithm Cross Validation Accuracy

Integrated sources	5-fold Accuracy					Mean accuracy
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Two sources	0,71	0,64	0,7	0,73	0,76	0,71
Three sources	0,59	0,56	0,59	0,66	0,74	0,63
Four sources	0,75	0,7	0,76	0,76	0,78	**0,75**

Table 5. The Two-Level Algorithm Accuracy

Integrated sources	Parameters					5folds Accuracy	80 / 20	70 / 30
	C	gamma	kernel	Alpha	Max Iter.			
Two sources	1	0,01	poly	1,00E-05	2000	0,71	0,81	0,71
Three sources	10	0,001	poly	1,00E-05	2000	0,63	0.84	0.72
Four sources	1	0,01	poly	1,00E-05	2000	0,75	**0,86**	0,77

7. Conclusions

This work built a relevant model for forecasting shares' directional movement from financial news articles. This is a unique research as it is the only paper predicting stock market in a parallel mode from four different web journals. We reached an accuracy of 86% which is a 26% improvement over a single data source model and even better over most similar works that typically have an accuracy situated between 50% and 70% [17]. For Training and testing the produced model, we scraped four financial Moroccan web journals. The number of extracted articles about any of the Moroccan market companies is low to feed any machine learning algorithm. Hence, we considered all articles corresponding to Casablanca Exchange stock market companies. This can have an impact on the performance of the model and accuracy can even be better if all articles are linked to a single company as demonstrated by [18]. The top level of the designed model inspected the correlation between vectorized articles and shares' prices movement. The vectorization process was based on the bag of words technic which leads generally to sparsity and high dimensionality problems. The use of a financial dictionary could decrease considerably the sparsity and dimension of features space. Bag of words abstract the semantic load of words and the context issues. So, the integration of a sentiment analysis component in this model is mandatory and could enhance the performance of this algorithm. This model is naturally not limited to news articles but can also be fed with tweets and social media posts; thus, a multi-type data source can be evaluated in a subsequent work.

References

1. Fama, E. (1964). *The behavior of stock Market Prices*. Graduate School of business University of Chicago.
2. Poterba, J.M., & Summers, L.H. (1988). Mean reversion in stock prices: Evidence and implications. *Journal of financial economics*, *22*(1), 27-59.
3. Charest, G. (1978). Dividend information, stock returns and market efficiency-II. *Journal of Financial Economics*, *6*(2-3), 297-330.
4. Charest, G. (1978). Dividend information, stock returns and market efficiency-II. *Journal of Financial Economics*, *6*(2-3), 297-330.
5. Hicham, E.B., & Salah-Ddine, K. (2020). Financial news analysis for moroccan stock trend predictions, Test Engineering Management., *82*(1-2), 1712–1717.
6. Gupta, E., Preetibedi, P.O.O.N.A.M.L.A.K.R.A., & Mlakra, P. (2014). Efficient Market Hypothesis V/S Behavioural Finance. *IOSR Journal of Business and Management*, *16*(4), 56-60.
7. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259-268.
8. Nelson, D.M., Pereira, A.C., & De Oliveira, R.A. (2017). Stock market's price movement prediction with LSTM neural networks. *In International joint conference on neural networks (IJCNN)* 1419-1426.
9. Kambeu, E. (2019). Trading volume as a predictor of market movement: An application of Logistic regression in the R environment. *International Journal of Finance & Banking Studies*, *8*(2), 57-69.
10. Bhuvaneshwari, C., & Beena, R. (2020). Stock Market Forecasting from Multi-Source Data using Tolerance Based Multi-Agent Deep Reinforcement Learning. *International Journal of Engineering and Advanced Technology (IJEAT)*. *9*(3), 3492–3499.
11. Ishijima, H., Kazumi, T., & Maeda, A. (2015). Sentiment analysis for the Japanese stock market. *Global Business and Economics Review*, *17*(3), 237-255.
12. Waduge, N., & Ganegoda, U. (2018). Forecasting Stock Price of a Company Considering Macroeconomic Effect from News Events, *3rd International Conference on Information Technology Research (ICITR)*, 3, 1–5.
13. Zhang, X., Qu, S., Huang, J., Fang, B., & Yu, P. (2018). Stock market prediction via multi-source multiple instance learning. *IEEE Access*, *6*, 50720-50728.

14. Li, X., Wang, C., Dong, J., Wang, F., Deng, X., & Zhu, S. (2011). Improving stock market prediction by integrating both market news and stock prices. In *International conference on database and expert systems applications*, Springer, Berlin, Heidelberg, 279-293.
15. Nayak, A., Pai, M.M.M., & Pai, R.M. (2016). Prediction Models for Indian Stock Market. *Procedia Computer Science*, *89*, 441–449.
16. Kaushal, A., & Chaudhary, P. (2017). News and events aware stock price forecasting technique. *In International Conference on Big Data, IoT and Data Science (BID)*, 8-13.
17. Khadjeh, N.A., Aghabozorgi, S., Ying, W.T., & Ngo, D.C.L. Text mining of news-headlines for Forex.
18. Aase, K.G., & Öztürk, P. (2011). Text Mining of News Articles for Stock Price Predictions, Department of Computing Information Science, *82*.