Stock Market Random Forest-Text Mining System Mining Critical Indicators of Stock Market Movements

1st Mazen Nabil Elagamy
Computer Engineering
Arab Academy for Science and Technology and Maritime Transport
Alexandria, Egypt
m_elagami@hotmail.com

2nd Clare Stanier
School of Computing and Digital Technologies
Staffordshire University
Stock-On-Trent, United Kingdom
c.stanier@staffs.ac.uk

3rd Bernadette Sharp
School of Computing and Digital Technologies
Staffordshire University
Stock-On-Trent, United Kingdom
b.sharp@staffs.ac.uk

***Abstract*— Stock Market (SM) is believed to be a significant sector of a free market economy as it plays a crucial role in the growth of commerce and industry of a country. The increasing importance of SMs and their direct influence on economy were the main reasons for analysing SM movements. The need to determine early warning indicators for SM crisis has been the focus of study by many economists and politicians. Whilst most research into the identification of these critical indicators applied data mining to uncover hidden knowledge, very few attempted to adopt a text mining approach. This paper demonstrates how text mining combined with Random Forest algorithm can offer a novel approach to the extraction of critical indicators, and classification of related news articles. The findings of this study extend the current classification of critical indicators from three to eight classes; it also show that Random Forest can outperform other classifiers and produce high accuracy.**

Keywords*— *Knowledge Discovery, Text Mining, Natural Language Processing, Stock Market, Random Forest.

I. Introduction

Text mining has become an important research area and has been applied in diverse domains ranging from medical, business and finance. Typical text mining applications cover the area of customer relationship management, market analysis, fraud detection, and in particular, stock market crashes analysis, which is an important event of today's global economy. As more textual data and news articles are increasingly available in a digital format, the business and financial sectors have recognised the competitive intelligence contribution that text mining can offer them in addition to data mining.

The increasing importance of stock markets and their direct influence on the economy have motivated this research to investigate the critical indicators, which characterise stock markets movements which could help support important business and financial decision making. In 2009, a number of events led to the United Arab Emirates crisis; these include the global recession, the bursting of the Dubai property bubble, and the post Lehman shutdown of international capital markets hit simultaneously. Dubai witnessed a major slowdown in growth and strains in its banking system as a result of the global financial crisis, the decline in oil prices, and in particular the bursting of its property bubble [1]. Our study of the stock markets movements is based on a total of 544 financial news articles concerning Dubai's stock market, published in the period between 2008 till 2012. This textual data include articles published before the crisis and after the crisis within the period for recovery of Dubai's SM (Dubai's SM upturn).

Text mining is a multidisciplinary field, which involves information retrieval, natural language processing, machine learning and statistics. The focus of this paper is to demonstrate the contribution of text mining in investigating the critical indicators of stock market movements using Random Forest classifier as the modelling approach. Random Forest is an efficient and interpretable machine learning algorithm; it achieves accurate predictions for various types of datasets because it uses ensemble strategies and random sampling. The model interpretability and the prediction accuracy of Random Forest are very rare among most of the machine learning algorithms. Furthermore Random Forest is less responsive to outliers in training data and there is no need to prune the trees because the bootstrapping and ensemble scheme makes Random Forest capable of overcoming the problems of overfitting. So Random Forest has all the advantages of decision trees and it achieves better results most of the times due to its utilisations of bagging on samples, random subsets of variables and voting schemes [2], [3], [4] and [5].

This paper is organised as follows. Section one has introduced the background of our study. Section two reviews current approaches to the study of financial applications. Section three describes our text mining system approach, Stock Market Random Forest-Text Mining system, to discover the critical indicators associated with the 2009 Dubai stock market debt standstill. The last two sections discuss the findings and the results of our approach.

978-1-5386-4543-7/18/$31.00 ©2018 IEEE

II. Related Work

The application domain of this research is stock market also known as equity market or share market; it is the market where shares of public listed companies are issued and traded [6]. The prediction of stock markets movements is significant for economical researchers from more than one perspective. Empirically, studying stock markets movements reveal information about stock markets' driving factors. From a theoretical point of view, this can be viewed as assessments of existing asset pricing theories. Hence, there are extensive studies in financial economics, which addressed this issue [7].

The use of data mining techniques to analyse stock markets has been extensively studied using structured data like past prices, historical earnings, or dividends, whereas text mining approaches are comparatively rare due to the difficulty of extracting relevant information from unstructured data. [8] stated that stocks behave randomly, and [9] and [10] explained that the application of data mining to the analysis of stock market data using current approaches might not be sufficient to model and justify any random behaviour of the market based only on quantitative data such as the values of stocks and historical market prices. They suggested that, if researchers focus on the impact of unquantifiable events on the market, which can be extracted from related news articles, they might be able to justify the random behaviour of the market and to enhance the analysis performance. [11] explained that there are huge amounts of free news and financial data, which are believed to contain rich information known as "alpha". Alpha is considered to be valuable, non-trivial and rich information embedded in textual data, which can be very useful for the purpose of analysis. The hypothesis of his research is that text mining approaches can enhance the performance of current trading systems' strategies if the "alpha" embedded in financial news is used to support the prediction of stock market share price movement directions [12].

The computational processing of unstructured textual data is considered to be one of the main reasons for the limited text mining development in stock market research. Textual data such as news, reports and economical articles are qualitative data, which must be converted to numeric form before many computational systems can process them. However, they are an important source of information about stock market and their analysis may provide a better understanding of random behaviour of the market, which is difficult to explain by focusing solely on numerical data [9]. Furthermore, many financial managers have been unable to fully exploit this valuable information, because it is implicit in the data and not easy to discern [13]. Textual information is complex and rich, whilst tables with financial data indicate how well a company has achieved, the linguistic structure and written style of the text may reveal more about its strategy and future performance [14]. The key issue is the necessity to use the user's specification to label historical documents for training and classifying.

The use of textual data, such as news, financial reports and economical articles, relies heavily on human analysis in order to achieve a better analysis of stock market price movements. It is important to note that relying solely on the analysis of statistical data has some limitations. The importance of news events can only be evaluated at a later time, and experts may have different opinions and interpretations of the events. Also the lack of sufficient and clear information about relationships between decision variables and outcomes always make experts and investors lapse into making relatively less rational decisions in financial market. This problem becomes worse when decision makers are confronted with large amounts of information [15]. Unlike numerical and fixed field data, textual date cannot be analysed by standard statistical data mining method [16]. Even though text mining is expected to play an important role in the financial sector, to the best of our knowledge the analysis of market behaviour and the need for designing strategies to cope with the movements of stock markets is still a relatively new field, which requires to be investigated [10].

Consequently, many researchers have explored the field of text mining to examine the effectiveness of text analysis for stock price movement prediction. [17] produced a text mining prediction system to forecasts companies' stock price changes (down, stay or up) influenced by financial events documents. Their results showed that textual analysis enhanced the prediction accuracy around 10% over a powerful baseline, which only deploys data mining techniques to analyse numeric data. [18] employed text mining to analyse financial news articles and reports in conjunction with time-series market data in order to explain the causes for poor performance or a sudden upturn in the market. They proposed a text mining system, which analyses financial news related to the Indian stock market in order to identify the major events, which have impact on the stock market and to design strategies for predicting the market. The events have been studied using Latent Dirichlet Allocation (LDA) based on topic extraction mechanism. The study carried out by [10] reveals that automatic text classification techniques are commonly used in analysing incoming news, and in some cases researchers make use of historical market prices data related to stock price to improve the accuracy of their prediction, thus combining data and text mining algorithms. Such predictive systems consist of three main components: classifier input generation, classification and finally news labelling. [19] propose a unified latent space model to characterise the "co-movements" between stock prices and news articles and to predict the closing stock prices on the same day; their algorithm is based on the analysis of daily articles from Wall Street journal. [20] predict the stock market based on textual information from user-generated micro-blogs using the latent space model to correlate the movements of both stock prices and social media content. [21] apply natural language processing to analyse economical news articles of a media company to categorise and extract the sentiments and opinions expressed by the writers. Their aim is to identify the correlation between news and stock market fluctuations. [22] adopted a linguistic based text mining approach demonstrating how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts. [23] examined a predictive machine learning approach to analyse financial news articles and stock quotes covering the S&P 500 stock market index during a five weeks period using a set of linguistic textual representations, including bag of words, noun phrases, and named entities approaches to estimate a discrete stock price

twenty minutes after a news article was released. Using Support Vector Machine (SVM) derivative tailored to discrete numeric prediction and models they showed that their model had the best performance in closeness to the actual future stock price and the highest return using a simulated trading engine. They have also concluded that a proper noun scheme performs better than bag of words in their metrics. [13] discussed various techniques (e.g. typical price, relative strength index and moving average) to predict whether future closing stock price will increase or decrease and to investigate various global events and their influence on predicting stock markets. [10] considered three market aspects, such as input data, predictive goal and prediction horizon, to predict the price and volatility of the market based on the new content. Using machine learning techniques they labeled the news and classified them to investigate the impact of financial news on stock market prediction. Similarly [24] classified financial news articles into positive or negative according to their effects on stock price based on price changes to label the articles and using support vector machine. [25] summarised studies, which are concerned with weighting text for predicting stocks price movements. In addition, they also reviewed the performance of various text mining methods applied using different text sources and they showed that most textual sources used by text mining researchers for market prediction include financial journals and news such the Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg and even Yahoo Finance, and often the analysis is focused on the news text or the news headlines. Consequently, [26] examined the use of textual data produced from users' micro-blogs in Tweeter to predict the stock market. They were able to find a correlation between the movement of stock prices and the social media content through the usage of the latent space model proposed by [19]. Their study did not evaluate sentiment of the social media data, whereas [27] proposed a sentiment analysis system based on summarisation to determine the polarity (positive or negative) of news articles from the Wall Street Journal and financial market data from the NASDAQ aimed at predicting the stock market. In addition, [28] constructed a predictive model to predict stock market future trends. Their model used sentiment analysis of multiple types of financial news and historical stock prices, which led to the achievement of prediction accuracy up to 89.80%.

III. Stock Market Radnom Forest-Text Mining (SMRF-TM)

This study has developed a text mining system, SMRF-TM, consisting of two stages: (i) applying natural language processing approach to analyse the textual data related to the 2009 Dubai stock market debt standstill in order to extract its critical indicator features, which can contribute to the prediction of abnormal stock movements, and (ii) analysing and classifying semantically these extracted critical indicators using Random Forest classifier and classifying the related news articles based on these relations (Fig. 1). The textual dataset selected to test our approach was obtained through a formal subscription in the official web site of the Financial Times. A total of 544 financial news articles concerning Dubai's stock market, published in the period between 2008 till 2012, were retrieved. This specific period was chosen so that it includes articles published before the crisis and after the crisis within the period for recovery of Dubai's stock market (Dubai's SM upturn). These 544 articles, which have around 1031006 total number of words are used for training and testing and served the basis to investigate the validity of the proposed SMRF-TM approach. The retrieved data is used to quantitatively validate and analyse the proposed approach using k-fold cross validation and text mining techniques such as, term frequency-inverse document frequency, random forest, and expectation maximisation in order to identify the critical indicators, which can seriously affect the prediction performance of stock market movements. Then a qualitative validation of the results yielded was carried out using financial experts.

The natural language processing stage which is described in [29] and [30], include three tasks, namely lexical analysis (e.g. tokenisation, removal of stop words, word stemming), syntactic analysis to generate unigrams and bigrams likely to be potential indicator features, and extracting those features, which capture the stock market movements. A total of 15,276 unigrams features and 103,506 bigrams features are generated. A vector-space model is used to capture the relevant extracted features for each article/document within our data. We can represent each document as a vector *(v)* in the *(t)* dimensional space if we have a set of *(d)* documents (i.e. articles) and a set of *(t)* terms. The features extraction stage produces a two-dimensional vector space where the rows represented the articles and the columns represented the features, and the cells capture the TF/IDF value for each feature. SMRF-TM applies TF/IDF to remove all the tokens with a threshold less than a set of different values and the results yielded from all these values were compared to check for the best threshold to be set.

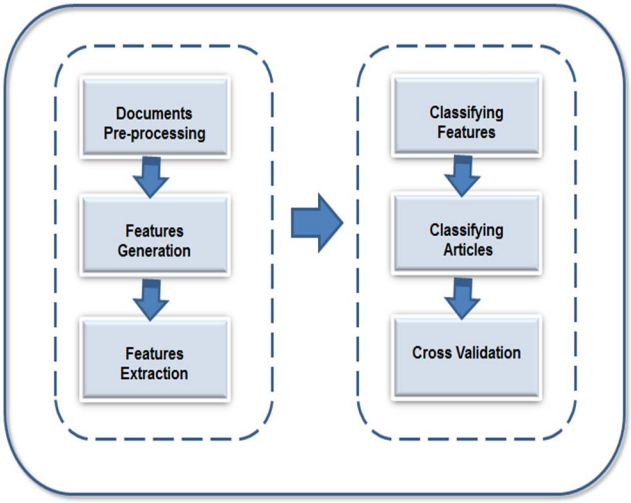

Fig. 1. The two stages of the SMRF-TM system.

The second stage involves the analysis and classification of these extracted features into appropriate semantic classes followed by cross validation. The features are first classified into critical down, down, neutral, up, critical up, and then further grouped into economic, social and political entities. This approach extends the three current main classes (down, neutral and up) adopted by [27], [31], [10], [32] and [24]. To discover the hidden knowledge and relations between our extracted features we have applied the Random Forest

algorithm, which is a machine learning algorithm, developed by [5]. Random Forest is a supervised classification algorithm, able to classify large amounts of data with high accuracy. It is a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest [33]. The basic principle is that a group of "weak learners" can come together to form a "strong learner".

An ensemble of 10 random, individual and unpruned trees is generated through the application of Random Forest in the SMRF-TM system. Each individual tree is constructed using the algorithm shown in Fig. 2. The relationships between the unigrams and bigrams features as classified by Random Forest are illustrated in Fig. 3 and Fig. 4.

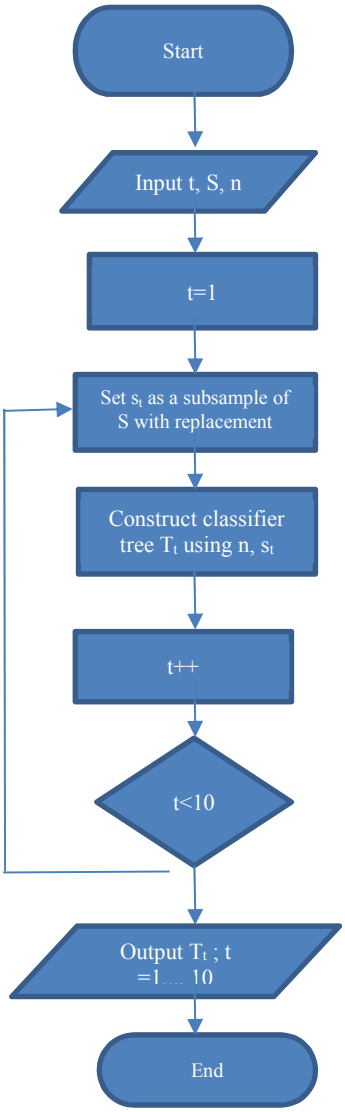

Fig. 2. Flowchart of the algorithm used to construct each individual tree in the Random Forest.

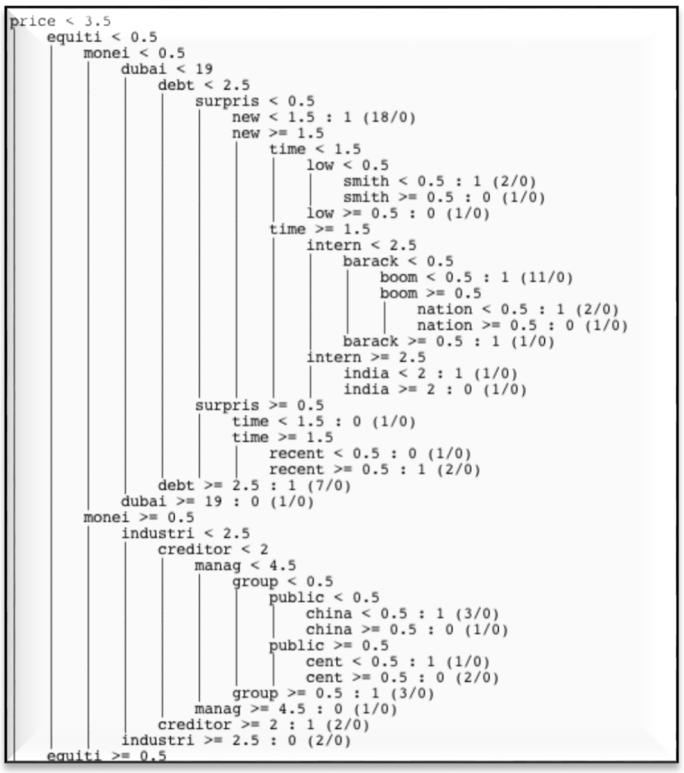

Fig. 3. Sample of how Random Forest discovers the relationships between unigram features in the SMRF-TM approach.

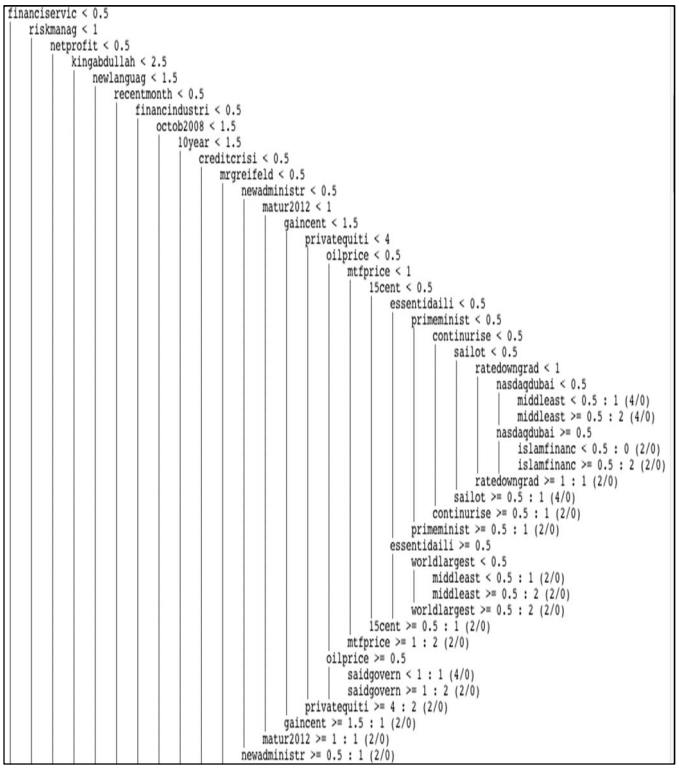

Fig. 4. Sample of how Random Forest discovers the relationships between bigram features in the SMRF-TM approach.

A cross validation with different folds (5, 10, 20, 30, 40, and 50) were used in order to validate the results and check which classifier yields the best classification performance. For example the 10-folds cross validation used 9/10 of the data for training the algorithm and 1/10 of the data for testing; this process is repeated 10 times after shuffling the data, each time. Tables 1, 2, 3, 4 and 5 show examples of the Random Forest classification results of the unigram features. Tables 6, 7, 8, 9 and 10 show examples of the Random Forest classification results of the bigram features.

Table 1. Examples of unigram features classified into the class "down".

Need	Debt	Conflict	Nakheel	Recess
Regime	Govern	Asset	Compliant	Destroy
Sale	Rate	Wage	Inflation	Risk
Elect	Fund	Report	Problem	Grow
Credit	Shortage	Downgrade	Emergency	Loan

Table 2. Examples of unigram features classified into the class "up".

Bond	Investor	Market	Transfer	Share
Trade	Finance	Growth	Develop	Fund
Federal	Construct	Secure	Consolidate	Manage
Legal	Propos	Creditor	Restructure	Corporate
Sukuk	Afford	Repaid	Sponsorship	Collaborate

Table 3. Examples of unigram features classified into the class "neutral".

Time	South	Nation	Year	Rate
Bond	Global	Bank	Term	Like
Metal	Firm	Europe	Paper	Gulf
World	Sale	Dubai	Saudi	Cent
Africa	London	State	Mine	Research

Table 4. Examples of unigram features classified into the class "critical down".

Decrease	Burst	Response	Crisis	Problem
Weak	Tighten	Squeeze	Quit	Fundament
Inflate	Limit	Trap	Risk	Withdraw
Hit	Boom	Impact	Reduce	Negative
Volatile	Gradual	Spare	Low	Shortage

Table 5. Examples of unigram features classified into the class "critical up".

Trade	Recover	Fund	Predict	Launch
Develop	Strengthen	Guard	Good	Deal
Stronger	Concern	Excess	Invest	Ultimate
Volume	Supreme	Accept	Higher	Win
Increase	Rate	Better	Rise	High

Table 6. Examples of bigram features classified into the class "down".

EmergeMarket	ExchangeRate	CentralBank
Weak Market	InflationExpected	MediumTerm
CompetitionTougher	IllegalImmigration	EstateAgent
PublicDebt	DebtRisk	DebtOffice
LoanGiven	CourtLaw	SmallNumber
ExistShareholder	ShortTerm	JobLoss

Table 7. Examples of bigram features classified into the class "up".

ConvertBond	MidMarket	RowPrice
ForeignInvestor	AssetManage	InvestBank
FianncialSupport	ManageTeam	LegalService
FinancialCentre	ShareholdActive	ProductRegion
FixFee	ShareTrade	HigherSalary

Table 8. Examples of bigram features classified into the class "neutral".

SouthKorea	PensionFund	GoldPrice
GoldMarket	MiddleEast	CreditMarket
WorldGold	BankReserve	MonthPeriod
IranTrade	BritishDiplomat	WesternBank
ArabianAutomobil	StandardBank	GulfCompany

Table 9. Examples of bigram features classified into the class "critical down".

UnemploymentGrowth	WeakEconomy	RaiseDebt
InflationRisk	BiggestRisk	DepressGrowth
PublicFinance	ConsumSpend	SlowRecover
CreditCondition	HigherTax	SharpFall tum
CapitalReduce	OverSell	BigDebt

Table 10. Examples of bigram features classified into the class "critical up".

DevelopEconomy	RiseMarket	RiseRate
GlobalBond	AveragUp	StrongEconomy
ImproveMarket	MarketRecovery	StockGrowth
DebtPaid	GlobalInvest	StrongGovern
Overbuy	HelpEconomy	GoForward
HighInvest	CapitalTrust	HighInterest

IV. Discussion

In summary, using the unigram features the Random Forest classifier has correctly classified 535 articles out of 544 of the dataset corpus resulting in 98.34% classification accuracy. The total number of the incorrectly classified articles using the Random Forest classifier is 9 articles out of 544, which are distributed as follows: two articles out of 134 for the neutral class resulting in 132 articles true positive, two articles false positive and two articles false negative. The down class has two incorrectly classified articles out of 184 yielding 182 articles true positive, two articles false positive and two articles false negative. The up class has four incorrectly classified articles out of 138 for producing 134 articles true positive, five articles false positive and four articles false negative. The critical down class has 54 articles correctly classified so it has 54 articles true positive, zero articles false positive and zero articles false negative. Finally, the critical up class has one incorrectly classified article out of 34, which means that it has 33 articles true positive, zero article false positive and one article false negative.

Using the bigram features the Random Forest classifier has correctly classified 538 articles out of 544 in the dataset corpus resulting in 98.89% classification accuracy. The total number of the incorrectly classified articles using the Random Forest classifier is six articles out of 544, which are distributed as follows: the neutral class has two incorrectly classified articles out of 134, so it has 132 articles true positive, four articles false positive and two articles false negative. The down class has two incorrectly classified articles out of 184, which means it has 182 articles true positive, zero articles false positive and two articles false negative. The up class has two incorrectly classified articles out of 138, 136 articles true positive, two articles false positive and two articles false negative. Regarding the critical down and the critical up classes they did not have any misclassified articles, hence, they have zero articles false positive and zero articles false negative.

In addition we have applied the Expectation Maximisation (EM) clustering algorithm which iteratively refines initial mixture model parameter estimates to better fit the data and terminates at a locally optimal solution [34]. In SMRF-TM, EM was applied to cluster the classified news articles according to their semantic meanings in one of the three clusters: economical, social/geographical or political facet. The application of EM clustering technique resulted in clustering 93% of the news articles to the economical cluster, 5% to the social cluster and 2% to the political cluster. Such results are considered reasonable given the source of our dataset is retrieved from the Financial Times.

To the best of our knowledge, Random Forest has not been applied using textual financial data to analyse stock market movements. Consequently, we can only compare the classification performance of other classifiers to the classification performance of Random Forest. The classification results produced by Random Forest were compared against the following ensemble of classifiers: Rotation Forest, Bagging, J48, Bayes Net, Decision Table, and Decision Stump. In order to achieve a better composite global model than using a single model to improve accuracy and thus obtain reliable estimates or decisions, the basic concept behind validation is to combine a set of models where each of these algorithms classifies the same original data.

Table 11 and Table 12 summarise the best results obtained using unigram and bigram features for each algorithm, expressed in terms of precision and recall measures defined below. The F1 score is a combined metric, which represents a balanced harmonic mean of precision and recall metrics and it is sometimes referred to as effectiveness measure. The F1 is considered as a standard performance index commonly used in machine learning to evaluate the classification performance in precision and recall space. The F1 score is able to evaluate the tested algorithms and produce an objective comparison of two or more algorithms and that is why we used it in our comparative study of the tested classifiers. Random Forest achieved the best performance using bigram features with 98.89% accuracy with 40 folds and decision stump achieved the worst performance overall with 43.75% accuracy measure (see Table 11 and Table 12).

$$Precision\ (PR) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Positive\ (FP)}$$

$$Recall\ (RC) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Negative\ (FN)}$$

$$Accuracy = \frac{Total\ number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances} * 100$$

$$F1 = \frac{2\ *\ PR * RC}{PR\ +\ RC}$$

Table 11. Summary of the best performance results of the 7 classifiers using unigram features.

Classifier	Cross Validation (folds)	Accuracy (%)	Precision (critical down class)	Recall (critical down class)	Precision (down class)	Recall (down class)	Precision (neutral class)	Recall (neutral class)	Precision (up class)	Recall (up class)	Precision (critical up class)	Recall (critical up class)
RF	40	98.34	1.00	1.00	0.99	0.99	0.99	0.99	0.96	0.97	1.00	0.97
Rotation Forest	40	92.83	1.00	1.00	0.95	0.91	0.89	0.93	0.90	0.92	0.97	0.91
Bagging	50	87.68	1.00	1.00	0.85	0.90	0.95	0.83	0.83	0.89	0.80	0.71
J48	30	84.00	0.96	1.00	0.84	0.84	0.84	0.84	0.86	0.78	0.63	0.79
Bayes Net	50	71.13	1.00	1.00	0.65	0.75	0.96	0.40	0.66	0.94	0.48	0.35
Decision Table	20	70.77	1.00	0.91	0.70	0.71	0.58	0.73	0.78	0.67	0.78	0.41
Decision Stump	5-10-20-30-40-50	43.75	1.00	1.00	0.38	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 12. Summary of the best performance results of the 7 classifiers using bigram features.

Classifier	Cross Validation (folds)	Accuracy (%)	Precision (critical down class)	Recall (critical down class)	Precision (down class)	Recall (down class)	Precision (neutral class)	Recall (neutral class)	Precision (up class)	Recall (up class)	Precision (critical up class)	Recall (critical up class)
RF	40	98.89	1.00	1.00	1.00	0.99	0.97	0.99	0.99	0.99	1.00	1.00
	50	98.89	1.00	1.00	0.99	1.00	0.99	0.97	0.99	0.99	1.00	1.00
Rotation Forest	30	86.21	1.00	1.00	0.85	0.89	0.88	0.85	0.82	0.82	0.83	0.74
	40	86.21	1.00	1.00	0.86	0.88	0.87	0.84	0.81	0.85	0.86	0.71
Bagging	30	81.98	0.98	1.00	0.77	0.86	0.87	0.75	0.80	0.80	0.76	0.65
J48	50	81.25	0.96	1.00	0.84	0.80	0.81	0.74	0.75	0.83	0.71	0.79
Bayes Net	50	75.36	1.00	1.00	0.63	0.91	0.98	0.59	0.77	0.77	0.67	0.12
Decision Table	50	73.52	1.00	0.91	0.65	0.83	0.77	0.63	0.76	0.68	0.70	0.56
Decision Stump	5-10-20-30-40-50	43.75	0.96	1.00	0.38	1.00	0.00	0.00	0.00	0.00	0.00	0.00

V. CONCLUSION

The stock market is a significant sector of a country's economy and represents a crucial role in the growth of their commerce and industry. Hence, discovering efficient ways to analyse and visualise stock market data is considered a significant issue in modern finance. [8] explained that stock markets behave randomly; consequently the application of data mining to the analysis of stock market data may not be sufficient to model and justify any random behaviour of the market. Given the huge amounts of free news and financial data, it is important to search for novel computational theories and tools to analyse and extract valuable hidden insights from this explosive growth of digital textual data. This study is an attempt at addressing this issue by extracting the critical indicators from unstructured yet valuable source of information, discovering the relationships between these indicators and classifying the news articles based on these relations. Random Forest classifier has a number of strengths, which makes it worthwhile to further investigate and apply to analysis stock markets articles.

This study has extended current approaches in classifying indicators from three to eight classes using Random Forest algorithm. It has also demonstrated that Random Forest can outperform the other classifiers (i.e. Rotation Forest, Bagging, J48, Bayes Net, Decision Table, and Decision Stump) and achieve the best accuracy in classifying the news articles based on bigram features. It is proposed to apply this approach to other sources of financial news articles such as Bloomberg, Reuters, Wall Street Journal, and Economic Times and other stock market crises in order to refine the discovery of critical indicators, which can be used to enhance the classification performance of the news articles leading to a better prediction of stock markets movements.

REFERENCES

[1] IMF United Arab Emirates 2009 Article IV Consultation — Staff Report; Public Information Notice; and Statement by the Executive Director for United Arab Emirates, IMF Country Report No. 10/42, February (2010)

[2] Kumar, N. and Khatri, S.: Implementing WEKA for medical data classification and early disease prediction. In: *3rd International Conference on Computational Intelligence & Communication Technology (CICT),* IEEE, (2017) 1-6

[3] Horning, N.: Introduction to decision trees and random forests. *American Museum of Natural History's*, (2013)

[4] Qi, Y. (2011). Random Forest for Bioinformatics. www.cs.cmu.edu/~qyj/papersA08/11-rfbook.pdf, (accessed on September 2017).

[5] Breiman, L.: Random forests. *Machine Learning*, 45(1): (2001) 5-32

[6] Cheema, A., Vora, A., Jain, C., Kataria, P., Shah, R. and Wagh, S.: Stock Forecasters, (2008)

[7] Pönkä, H.: Predicting the direction of US stock markets using industry returns. *Empirical Economics*, *52*(4), (2017) 1451-1480

[8] Patel, J., Shah, S., Thakkar, P. and Kotecha, K.: Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications 42(4), (2015) 2162-2172

[9] Schumaker, R. P., Zhang, Y., Huang, C. N. and Chen, H.: Evaluating sentiment in financial news articles. Decision Support Systems 53(3), (2012) 458-464

[10] Nikfarjam, A., Emadzadeh, E. and Muthaiyah, S.: Text mining approaches for stock market prediction. In: The 2nd International Conference on Computer and Automation Engineering, ICCAE, vol. 4, IEEE, (2010) 256-260

[11] Drury, B.: A Text Mining System for Evaluating the Stock Market's Response To News. Doctoral dissertation in Computer science, University of Porto (2013)

[12] Kumar, B.S. and Ravi, V.: A survey of the applications of text mining in financial domain. Knowledge-Based Systems 114, (2016) 128-147

[13] Kannan, K.S., Sekar, P.S., Sathik, M.M. and Arumugam, P.: Financial Stock Market Forecast using Data Mining Techniques, *International MultiConference of Engineers and computer scientists.* Hong Kong,1, (2010)

[14] Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H. and Visa, A.: Combining data and text mining techniques for analysing financial reports. Intelligent systems in accounting, finance and management 12(1), (2004) 29-41

[15] Wang, S., Xu, K., Liu, L., Fang, B., Liao, S. and Wang, H.: An Ontology Based Framework for Mining Dependence

Relationships Between News and Financial Instruments, *Expert Systems with Applications,* 38**,** (2011) 12044-12050

[16] Nasukawa, T. and Nagano, T.: Text Analysis and Knowledge Mining System, *IBM Systems Journal,* 40,4, (2001) 967-984

[17] Lee, H., Surdeanu, M., Maccartney, B. and Jurafsky, D.: On the Importance of Text Analysis for Stock Price Prediction. In *LREC*, (2014) 1170-1175

[18] Mahajan, A., Dey, L. and Haque, S. M.: Mining Financial News for Maior Events and Their Impacts on the Market, In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. WI-IAT.* IEEE Computer Society, Sydney, Volume (1), (2008) 423-426

[19] Ming, F., Wong, F., Liu, Z. and Chiang, M.: Stock market prediction from WSJ: text mining via sparse matrix factorization. In: IEEE International Conference on Data Mining (ICDM), (2014) 430-439

[20] Sun, A., Lachanski, M. and Fabozzi, F.J.: Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis 48, (2016) 272-281

[21] Kim, Y., Jeong, S.R. and Ghani, I.: Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl. 6(1), (2014)

[22] Ali, M.M.Z. and Theodoulidis, B.: Analyzing Stock Market Fraud Cases Using a Linguistics- Based Text Mining Approach. In: WaSABi-FEOSW@ ESWC, (2014)

[23] Schumaker, R.P. and Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS) 27(2), 12 (2009)

[24] Kaya, M.Y. and Karsligil, M.E.: Stock price prediction using financial news articles. In: 2nd IEEE International Conference on Information and Financial Engineering (ICIFE), (2010) 478-482

[25] Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L.: Text mining for market prediction: A systematic review. Expert Systems with Applications 41(16), (2014) 7653-7670

[26] Sun, A., Lachanski, M. and Fabozzi, F.J.: Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, 48, (2016) 272-281

[27] Sorto, M., Aasheim, C. and Wimmer, H.: Feeling The Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment. In: Proceedings of the Southern Association for Information Systems Conference 2017. St. Simons Island, GA, USA (2017)

[28] Khedr, A.E., Salama, S.E. and Yaseen, N.: Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. International Journal of Intelligent Systems and Applications (IJISA) 9(7), (2017) 22-30

[29] Elagamy, M.N., Stanier C. and Bernadette S.: Text Mining Approach to Analyse Stock Market Movement. In *the 3rd International Conference on Advanced Machine Learning Technologies and Applications* (AMLTA), Springer (2018)

[30] Elagamy, M.N.: Text Mining Approach to Analyse Critical Indicators of Stock Market Movements. Doctoral dissertation in Computer science, Staffordshire University, UK (2018)

[31] Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research 3(Mar), (2003) 1289-1305

[32] Martinez-Romo, J. and Araujo, L.: Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications 40(8), (2013) 2992-3000

[33] Myung, J., Yang, J.Y. and Lee, S.G.: Picachoo: a tool for customizable feature extraction utilizing characteristics of textual data. In: Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication, ACM, (2009) 650-655

[34] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P.: Supervised machine learning: A review of classification techniques. Informatica 31, (2007) 249-268