# Predicting S&P 500 based on its constituents and their social media derived sentiment

Rapheal Olaniyan[1,2], Daniel Stamate[1,2], Ida Pu[1,2], Alexander Zamyatin[1,3], Anna Vashkel[1], and Frederic Marechal[1]

[1] Data Science & Soft Computing Lab, London
[2] Computing Department, Goldsmiths, University of London
[3] Institute of Applied Mathematics and Computer Science, Tomsk State University

**Abstract.** Collective intelligence, represented as sentiment extracted from social media mining, is encountered in various applications. Numerous studies involving machine learning modelling have demonstrated that such sentiment information may or may not have predictive power on the stock market trend, depending on the application and the data used. This work proposes, for the first time, an approach to predicting S&P 500 based on the closing stock prices and sentiment data of the S&P 500 constituents. One of the significant complexities of our framework is due to the high dimensionality of the dataset to analyse, which is based on a large number of constituents and their sentiments, and their lagging. Another significant complexity is due to the fact that the relationship between the response and the explanatory variables is time-varying in the highly volatile stock market data, and it is difficult to capture. We propose a predictive modelling approach based on a methodology specifically designed to effectively address the above challenges and to devise efficient predictive models based on Jordan and Elman recurrent neural networks. We further propose a hybrid trading model that incorporates a technical analysis, and the application of machine learning and evolutionary optimisation techniques. We prove that our unprecedented and innovative constituent and sentiment based approach is efficient in predicting S&P 500, and thus may be used to maximise investment portfolios regardless of whether the market is bullish or bearish.

**Keywords:** Collective intelligence · Sentiment analysis · Stock market prediction · Feature selection · Feature clustering · PCA · Jordan and Elman Neural Networks · Evolutionary computing · Statistical tests · Granger causality.

## 1 Introduction

Stock market is considered to be highly volatile. With this inherent problem, developing an efficient predictive model using purely traditional stock data to capture its trends, is considered to be hard to achieve. On the other hand, behavioural finance relaxes the assumption that investors act rationally. It underlines the importance of sentiment contagion in investment. Since then, researchers have been focusing on the relationship between sentiment and the

stock market. For example, Shiller [18] and Sprenger et al. [19] observed that factors related to the field of behavioural finance influence the stock market as a result of psychological contagion which makes investors to overreact or underreact. They imply that investors have the tendency to react differently to new information which could be in the form of business news, online social networking blogs, and other forms of online expressions.

Observations from related research works sprang up interest in advancing the standard finance models to include collective intelligence information represented by sentiment extracted from social media mining, in the predictive model development, with the aim of enhancing the model reliability and efficiency. Yet in order to statistically validate this inclusion, one needs to consider the source of the sentiment, examine its statistical significance and the Granger causality [7] between the sentiment and the stock market variables by using appropriate approaches.

Gilbert and Karahalios [6] investigated the causal relationship between the stock market returns for S&P 500 and the sentiment based on a collection of LiveJournal blogs. The sentiment was considered as a proxy for public mood. Using a linear framework they showed that sentiment possesses predictive information on the stock market returns. An obviously arising question in such a context is: is the linear framework robust enough to examine the Granger causality between stock market returns and sentiment? Olaniyan et al. [15] presented their finding from the re-assessment of the work conducted by Gilbert and Karahalios [6]. They showed that the models in [6] presented flaws from a statistical point of view. [15] further investigated the causality direction between sentiment and the stock market returns using a non-parametric approach and showed that there is no line of Granger causality between the stock market returns and sentiment in the framework considered in [6].

The influence of sentiments on the stock market has been extensively studied and so are the asymmetric impacts of positive and negative news on the market. But little has been done in devising efficient predictive models that can help to maximise investment portfolios while taking into consideration the statistical relevance of sentiment, and the proposed work addresses this concern. The main aim of this research is to predict reliably the directions of the S&P 500 closing prices, by proposing a predictive modelling approach based on integrating and analysing data on S&P 500 index, its constituents, and sentiments on these constituents. Indeed, this study is the first work to use constituent sentiments and its closing stock prices containing over 800 variables (combined closing stock prices of the S&P 500 constituents and their respective sentiment data without taking into account lagging - which further increases data dimensionality in a n-fold fashion) to predict the stock market.

First we tackle the data high dimensionality challenge by devising and proposing a method of selecting variables by combining three steps based on variable clustering, PCA (Principal Component Analysis) [11], and finally on a modified version of the Best $GLM$ variable selection method developed by McLeod and Xu [14].

Then we propose an efficient predictive modelling approach based on Jordan and Elman recurrent neural network algorithms. To avoid the pitfall of time invariant relationship between the response and the explanatory variables in the highly volatile stock market data, our approach captures the dynamic of the explanatory variables set for every rolling window. This helps to incorporate the time-variant and dynamic relationship between the response and explanatory variables at every point of the rolling window using our variable selection technique mentioned above. Finally, we propose an efficient hybrid trading model that incorporates a technical analysis, and machine learning and evolutionary optimisation algorithms.

We prove that our constituent and sentiment based approach is efficient in predicting S&P 500, and thus may be used to maximise investment portfolios regardless of whether the market is bullish or bearish [4]. This study extends our previous recent work on XLE index constituents' social media based sentiment informing the index trend and volatility prediction [13].

The remainder of this paper is organized as follows. Section 2 presents our data pre-processing methodology, which is our proposed method for handling the data high-dimensionality challenge outlined above, for selecting the variables with predictive value. Section 3 elaborates on the results of the causality relationship between sentiment and the stock market returns using special techniques of Granger causality. Section 4 presents the predictive modelling approach that we propose based on machine learning techniques including Jordan and Elman recurrent Neural Network algorithms. Section 5 entails our proposed trading model that combines a technical analysis strategy and the estimated results from the machine learning framework to optimise investment portfolios with evolutionary optimisation techniques. Finally Section 6 discusses our findings and concludes the paper.

## 2 Stock data and sentiment information

In order to develop our approach to predicting the S&P 500 close prices, we rely on three main datasets which we integrate. The first dataset involves the collection of all the closing stock prices for the S&P 500 constituents, and is obtained directly from Yahoo Finance website. The second dataset is sentiment data for the constituents of the S&P 500 index, obtained from Quandle [5] [9]. And the third dataset contains the S&P 500 historical close prices and trading volume, obtained also from Yahoo Finance website.

[4] Bullish and bearish are terms used to characterize trends in the stock markets: if prices tend to move up, it is a bull market; if prices move down, it is a bear market.

[5] Quandl collect content of over 20 million news and blog sources real–time. They retain the relevant articles and extrapolate the sentiment. The sentiment score is generated via a proprietary algorithm that uses deep learning, coupled with a bag-of-words and n-grams approach. Negative sentiments are rated between -1 and -5, while positive sentiments are rated between 1 and 5.

All the data collected covered the period from 8th of February 2013 to 21st of January 2016. For S&P 500 and its constituents, the stock market return at time $t$ is defined as $R_t = log(SP_{t+1}) - log(SP_t)$, where $SP$ is the closing stock price. The stock market acceleration metric is obtained from the stock market return as $M_t = R_{t+1} - R_t$. Moreover, $V_t$ is expressed as the first difference of the logged trading volume. Finally, the sentiment acceleration metric is defined as $A_t = S_t - S_{t-1}$, where $S_t$ represents sentiment for each constituent of the S&P 500 at moment $t$.

By combining the three datasets, in all we have more than 800 initial variables to explore (not including lagged variables), which will lead to one of the challenges encountered in our framework in terms of high data dimensionality.

Figure 1 shows the data pre-processing process flow. It highlights all the processes undertaken to refine the data.

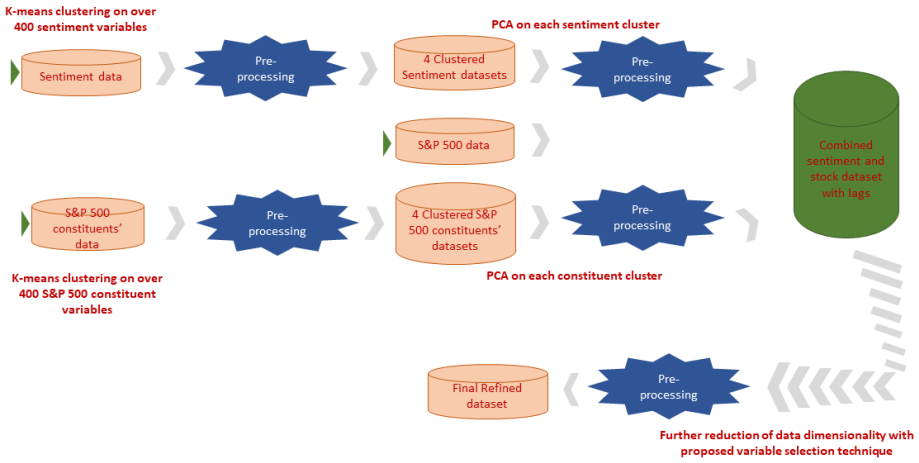


**Fig. 1.** Data pre-processing process flow that details the processes followed to tackle the complexity of high dimensionality dataset

To handle the high data dimensionality challenge, we propose an approach to reducing the number of dimensions, adapted to our framework, based on 3 steps, consisting consecutively of performing variable clustering, PCA, and by applying a variable selection method that we introduce here based on Best $GLM$ variable selection method developed by McLeod and Xu [14]. These steps are described in the following subsection.

### 2.1 Reducing data dimensionality

As mentioned, the prices and sentiments of the S&P 500 constituents are the variables of two of the datasets we dispose of initially. For analysis, it is important to classify each constituent into groups. Of course, classifying the constituents based on their respective industries would have been the easy way to group them since predefined information are readily available. Instead, we follow a more analytically rigorous approach in grouping the constituents based on pattern recognition and similarities in time series by using clustering.

In our case we use K-means clustering on each of the two sets of S&P 500 constituent and sentiment time series, respectively, in order to group the variables in clusters. On the other hand, as we intend to use a rolling window of 100 days for testing and 10 days for forecasting, we note here that clustering is therefore applied on each rolling window, by forming 4 clusters. We note that by exploring different numbers of clusters between 3 and 10, 4 was the optimal number of clusters which led to the best final outcomes in our framework. Due to the generic property of within cluster similarity, it is expected that variables in a cluster are more or less similar.

When it comes to reducing dimensionality of a numeric dataset, one of the most used methods is the well known Principal Component Analysis (PCA) [11]. Instead of applying PCA on all variables at once, we apply it on the groups of variables corresponding to each of the 4 clusters. Results from PCA show that dimensionality for the closing prices of S&P 500 constituents was reduced by 25% on average, and by 20% for sentiments. When we combine the principal components from both sentiments and closing prices we are still faced with a high number of dimensions of the combined dataset. By lagging the combined dataset up to 3 lags, its dimensionality increases to over 1200 variables, which keeps the intended predictive modelling at a challenging level computationally and from a predictive modelling point of view. This led us to propose a variable selection method to handle this complexity in our approach.

As random forest is a popular technique used in variable selection, it was our first choice method to consider in order to further reduce data dimensionality in this 3rd step of our approach. Interestingly, this solution performed very poorly on our dataset, judging by the poor goodness of fit with Adjusted R-Squared being below 0.3. We therefore proposed an alternative solution which is based on the modified best GLM method below developed by McLeod and Xu [14]. The latter selects the best subset of inputs for the $GLM$ family. Given output $Y$ on $n$ predictors $X_1, ..., X_n$, it is assumed that $Y$ can be predicted using just a subset $m < n$ predictors, $X_{i,1}, ..., X_{i,m}$. The aim is therefore to find the best subset of all the $2^n$ subsets based on some goodness-of-fit criterion. Consider a linear regression model with a number of $t$ observations, $(x_{i,1}, ..., x_{i,n}, y_i)$ where $i = 1, 2, ..., t$. This may be expressed as

$$M_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_n x_{i,n} + \epsilon_i \tag{1}$$

It is clear that when $n$ is large, building $2^n$ regressions becomes computationally too expensive, and even untractable in our case with $n > 1200$ predictors,

as mentioned above. As such, we modify McLeod and Xu's method of [14] as follows and we call the resulting method MBestGLM. First, the lagged dataset is divided into subsets whereby each subset contains 35 predictors and then the variable selection technique of [14] is applied on each subset with the intention of obtaining statistically significant predictors from each subset. The statistically significant predictors are then combined and the process of dividing the result into other subsets and applying the variable selection technique continues until the set of predictors can no longer be reduced. The regression result from this selected predictors produces a high adjusted R-Squared of over 0.65. The final dataset we obtain has an average number of predictors of 35. Indeed, from experiments we have seen that its number of predictors varies between 30 and 40 according to the instance of the rolling window on which the dataset is generated.

Overall, the dimensionality reduction process including the 3 steps of variable clustering, PCA, and the MBestGLM method we introduced above, are repeatedly applied on the rolling window as we work under the more general and thus more complex assumption of time-variant relationship between independent variables and return.

## 3 Sentiment's predictive information on S&P 500

As mentioned in the Introduction, it has been shown in a series of studies that sentiment variables help improve stock market prediction models ( [2], [6], [3] and [16]). In light of this it becomes imperative for us to investigate if the sentiment variables of S&P500 constituents included in our framework have some significant predictive power on this stock index.

In examining the relevance of sentiment variables we use two methods, the first based on linear models, and the second one, more general, based on non-linear non-parametric models, respectively. These are Granger causality statistical tests and are used to see if sentiment has predictive information on S&P 500 in our framework.

### 3.1 Granger causality test: the linear model

Using the linear model framework represented by the Granger causality statistical test [7], we examine the causal relationship between sentiment and stock market returns. According to [7] we write the general linear VAR models as:

$$Model1 : M_t = \alpha_1 + \Sigma_{i=1}^{3}\omega_{1i}M_{t-i} + \Sigma_{i=1}^{3}\beta_{1i}Stock_{t-i} + \epsilon_{1t} \quad (2)$$

$$Model2 : M_t = \alpha_2 + \Sigma_{i=1}^{3}\omega_{2i}M_{t-i} + \Sigma_{i=1}^{3}\beta_{2i}Stock_{t-i} + \Sigma_{i=1}^{3}\gamma_{2i}Sent_{t-i} + \epsilon_{2t} \quad (3)$$

where $M_t$ is the response variable which is the S&P 500 stock market return at time $t$, $M_{t-i}$ is the lagged S&P 500 market return with lag period of $i$, and $Stock$ and $Sent$ are variables generated by our 3-step dimensionality reduction process issued from stock components and sentiment variables respectively. These VAR

models $Model1$ and $Model2$ are used to examine if sentiment influences the stock market in our setting. As observed in the two equations, $Model1$ uses the lagged stock market return and the lagged stock market return principal components generated from the close prices of the S&P 500 constituents. In $Model2$ the lagged principal components, generated from sentiment variables related to the S&P 500 constituents, are added to the variables used in $Model1$. That is, $Model1$ does not contain sentiment variables while $Model2$ does. Sentiment variables would be considered to be influential if $Model2$ outperforms $Model1$ in prediction performance based on the adjusted R-squared metric. This is checked by using the standard Granger causality statistical test [7]. In particular we consider the hypothesis H0 that $Model2$ does not outperform $Model1$, and we reject it by obtaining a significant p-value.

Our results show that $Model2$, with the sentiment included in the analysis, outperforms $Model1$, based on the Granger causality $F$ statistics $F_{16,165} = 9.1438$, and the corresponding p-value $p_{Granger} < 0.0001$. Robust tests performed on the estimated residuals show that the residuals do not possess autocorrelation, are normally distributed and homoscedastic in variance (having p-values Ljung-Box $> 0.05$, and Shapiro-Wilk $> 0.05$), so the Granger causality test was applied correctly and its conclusion is valid. Thus sentiment has predictive information on S&P 500. In the next subsection we verify this conclusion with a more general non-parametric non-linear Granger causality test.

### 3.2 Granger causality test: the nonlinear model

Causality test from the linear model has already shown that sentiment variables have predictive power on the stock market. And the robust tests confirm that the results are not biased by the presence of autocorrelation or heteroscedasticity. Still, we examine the influence of sentimental information on the stock market using a non-linear non-parametric test which was originally proposed by Baek and Brock [1] and was later modified by Hiemstra and Jones [8].

Interestingly, the significant p-values from the nonlinear non-parametric technique (see [5] for detail explanation) displayed in Table 1 prove that sentiment has predictive power on the stock market.

**Table 1.** Nonlinear non-parametric Granger tests. $A$ and $M$ are the sentiment and stock market returns, respectively. $A => M$, for example, denotes the Granger causality test with direction from A to M, i.e. sentiment predicts stock market returns. Similarly, $M => A$ is a Granger causality test if stock market predicts sentiment.

| $Lx = Ly = 1$ | $p - value$ |
|---|---|
| $A => M$ | 0.0077 |
| $M => A$ | 0.0103 |

As a conclusion of this section, we can confidently state that the inclusion of sentiment variables does improve significantly stock market predictive models in terms of prediction performance, in our framework. Another interesting finding based on the significant p-value of $M => A$ in the nonlinear non-parametric Granger causality test, reveals that the stock market Granger-causes sentiment in this framework of S&P500 with its constituents and their sentiment.

## 4 Jordan and Elman neural network based approach to predicting S&P500 with sentiment

Linear and non linear models have been employed to assess the influence of sentiments on the stock market and results have shown the statistical significance of sentiments' influence on the stock market in our setting. A linear model has also been developed in the previous section (see $Model2$) to investigate if the future S&P 500 close prices can be predicted with sentiment.

This section evaluates the relative improvements to the linear model when we enhance our approach by using Recurrent Neural Networks algorithms, more specifically for Jordan and Elman networks. The backpropagation algorithm is one of the most popular techniques for training Neural Networks. It has been used in research works such as Collins et al. [4] which applied it to underwriting problems. Malliaris and Salchenberger [12] also applied backpropagation in estimating option prices. To determine the values for the parameters in the algorithms, the gradient descent technique is mostly employed Rumelhart and McClelland [17]. Multilayer, feed-forward, and recurrent Neural Networks such as Jordan and Elman Neural Networks which are used in this study, have become very popular.

As the datasets explored in our framework are highly dimensional, we rely on our variable selection methodology that we proposed in Subsection 2.1, to assist in selecting a reduced subset of variables based on S&P 500 index, its constituents and their sentiment, to implement a predictive modelling approach with Elman and Jordan Neural Network algorithms. That is, the same variable selection process used to obtain results from the estimated linear model in Section 3, is also used with our Neural Network models. It is important to note that in our approach we use a rolling window of 100 days for model development and fitting, and a rolling prediction period of 10 days. This choice was made based on several experiments we ran with our approach, whose details we don't include here due to lack of space. Knowing that the output of Neural Network models is sensitive to the values assigned to the parameters in the models (including the number of hidden layers, the number of their nodes, and the weights), with some computational efforts, fairly optimised Neural Network models have been generated. Since at each rolling window we may have different selection of the set of predictors, the values assigned to Neural Network parameters would therefore be expected to be different for each fairly optimal result.

As observed in Figures 2, the Elman Neural Network algorithm captures the stock market close price more accurately than the Jordan Neural Network

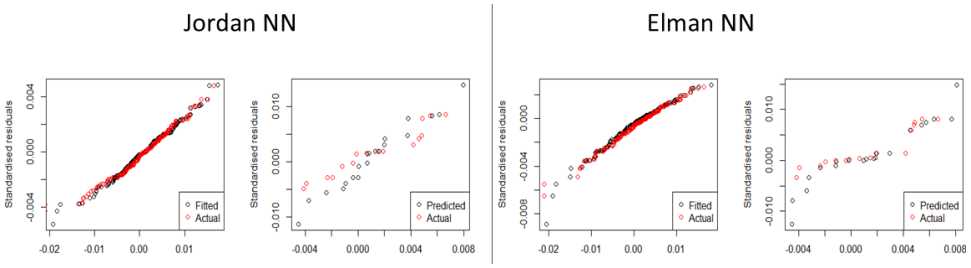


**Fig. 2.** Jordan Neural Network (Jordan NN) and Elman Neural Network (Elman NN): For each Neural Network, the left figure shows the fitted versus actual values, and the right figure shows the predicted versus actual values, all of which based on the dataset with rolling windows between 20/03/2014 and 12/08/2014

algorithm for both the fitted and predicted values. We conclude this section by mentioning that both Neural Network models outperformed the linear model developed in the previous section $Model2$ (details are not included due to lack of space).

## 5 Evolutionary optimised trading model

In the previous sections we have demonstrated that sentiments influence the stock market prices based on the results from linear and Neural Network frameworks. But with all the information we have so far, are we able to maximise investment portfolio by leveraging on the insightful information from our estimated models? We note that the information available still looks raw and therefore needs refining before we could make good use of it. In the process of refining the information, we resolve to introducing some stock market technical analysis and an evolutionary optimisation algorithm to our developed model. In doing so we propose the following strategies:

1. Active investment in *put* option with the expectation that price will fall in the future. The investor therefore profits from the fall in price. This helps to exploit bearish market.
2. Active investment in *call* option with the expectation that price will rise in the future. The investor therefore profits from the rise in price.
3. Hold position which implies that no investment should be made.
4. Passive investment refers to investment in stock market for a period of time without any optimal investment strategy.

Points 1 - 3 will be used to maximise investment portfolio under active investment and point 4 will be used to compare active and passive investment strategies.

The active investment strategies use the input from the estimated Neural Network models and also technical analysis data variable $K$, called the Chaikin

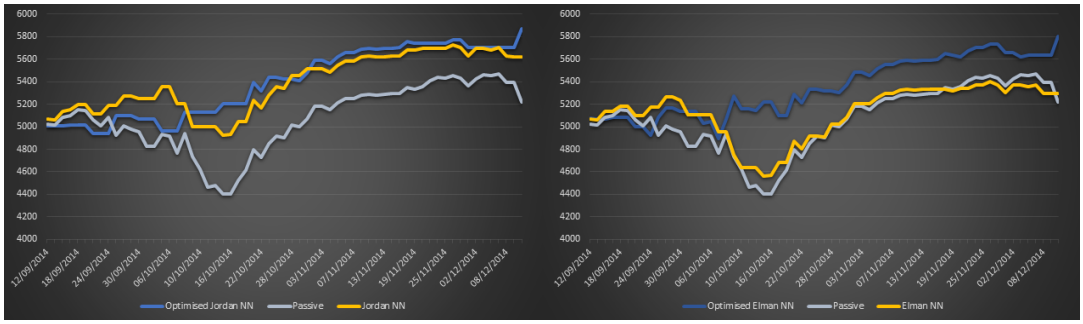


**Fig. 3.** The three investment portfolios are presented on two separate charts each related to Jordan Neural Networks (Jordan NN) on the left and Elman Neural Networks (Elman NN) on the right. The trends in Blue and Yellow present the optimised models and ordinary Neural Networks active investment portfolios respectively. The trend in Grey represents the passive investment portfolio.

Oscillator, which determines the position of the forces of demand and supply - see details on the calculation of the variable in [10]. To maximise the investment portfolio we employ an evolutionary optimisation algorithm. Given the objective investment function below:

$$\mathrm{f}(call, put) = \begin{cases} Invest_{n-1} + (Price_n - Price_{n-1}) & call \\ Invest_{n-1} + (Price_{n-1} - Price_n) & put \\ Invest_{n-1} & else \end{cases}$$

where
$call : Pred_n > a, \triangle K_{n-1} > b, \triangle K_{n-2} > c, \triangle K_{n-3} > d,$
$put : Pred_n < e, \triangle K_{n-1} < f, \triangle K_{n-2} < g, \triangle K_{n-3} < h,$ $Pred_n$ is the predicted value at day $n$, $\triangle K_n$ is the change in Chaikin Oscillator at day $n$, and $a, b, c, d, e, f, g, h$ are variables whose values must be determined. In order to maximise the objective investment function, we consider the following maximization problem:

$$\begin{aligned} \underset{a,b,c,d,e,f,g,h}{\text{maximise}} \quad & \mathrm{f}(call, put) \\ \text{subject to} \quad & -0.4 <= b, c, d, f, g, h <= 0.4 \end{aligned} \tag{4}$$

The evolutionary optimisation algorithm is then applied to Equation (4) in order to generate the values for $a, b, c, d, e, f, g$ , and $h$. The objective function is fairly optimised using just the first 35 days and the estimates obtained are kept constant to estimate portfolio values and trends for the next 100 days. Expectations regarding the relevance of this optimisation algorithms and technical analysis method are that trends obtained from the optimised models would be more stable than the ones that are not optimised. Also, we expect rising trends as these trends interpret to portfolio values. Decreasing trends would imply loss in investment. Looking at the results from Figure 3 it is clear that the optimised active models outperform the ordinary estimated machine learning models and

the passive portfolio. This conclusion is based on the fact that the trends in blue appear to be the most stable and fairly rising trends when compared with the trends from the ordinary estimated machine learning models. Even when persistent loss is reported in the passive portfolio in the period 07/10/2014 – 21/10/2014, trends from the optimised models appear fairly stable and rising. This is due to the fact that the optimised models take account of both bearish and bullish stock market using *put* and *call* options respectively.

## 6 Discussion and conclusion

This research work delivers its first novelty by the nature of the data explored, which at our best knowledge, was not considered by previous studies. For analysis purposes, our framework combined the closing prices of S&P 500 constituents and also their related sentiments which in total provides about 800 variables. This dimensionality challenge is n-fold increased due to lagging operation common with time series. To tackle the challenge of high dimensionality of the dataset in a computationally expensive prediction modelling approach that we proposed, a specially designed data pre-processing methodology was introduced. To the best of our knowledge, this is the first work to have used constituent sentiments and its closing stock prices (containing over 800 variables combining closing stock prices of the S&P 500 constituents and their respective sentiment data without lagging) in stock market predictive modelling.

With the rolling window of a 10-day predictions period and time-variant relationship between response variable and predictors - approach which involves obtaining a new set of predictors for every rolling window - the analysis' challenge became compounded. Random forest method failed to do a good predictor selection, as a first method of choice that we considered. As such we proposed a 3-step feature selection methodology involving the consecutive phases of variable clustering, PCA, and our own method of further feature selection that we call MBestGLM.

Having established the most significant variables in our proposed predictive modelling approach, and also justified the inclusion of sentiment in the approach as we proved its predictive value using Granger based methods, we develop models based on Recurrent Neural Network algorithms to predict the S&P 500 closing prices. However, this information per se is not sufficient enough to reliably predict the stock market trends and also maximise investment portfolios. As such, we enhanced our approach by proposing investment strategy models which make use of the generated estimates from the predictive models as input variables to bridge these gaps. Results show that our proposed model appears to be stable even when the stock market is bearish and other approaches are failing. The rationale is that the proposed model is engineered to perform using *put* and *call* options during bearish and bullish moments, respectively. This represents another novelty of our work.

We currently develop further work on exploring the extension of this approach and of the approach proposed in our recent work [13], for several stock market indices.

## References


1. Baek, E., Brock,W.: A general test for nonlinear Granger causality: bivariate model, Working paper. Iowa State University, 1992.
2. Baker,M.,Wurgler,J.: Investor sentiment in the stock market, Journal of Economics Perspectives, 21(2), 129–151.
3. Bollen,J., Mao,H., Zeng,X.: Twitter mood predicts the stock market, Journal of Computational Science, 2(1), 1–8, 2011.
4. Collins,E., Ghosh, S., Scofield,C.: An application of a multiple neural-network learning system to emulation of mortgage underwriting judgments,Proc, IEEE Int. Conf. on Neural Networks, 459–466, 1988.
5. Diks, C.,Panchenko,V.: A new statistic and practical guidelines for nonparametric Granger causality testing, Journal of Economic Dynamics and Control, 30(9-10), 1647–1669, 2006.
6. Gilbert,E., Karahalios,K.: Widespread worry and the stock market, In Proceedings of the 4th International Conference on Weblogs and Social Media, 58–65, 2010.
7. Granger, C. W. J.: Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica. 37 (3): 424–438, 1969.
8. Hiemstra,C., Jones,J.D.: Testing for linear and nonlinear Granger causality in the stock price–volume relation, Journal of Finance, 49, 1639–1664, 1994.
9. https://www.quandl.com/data/AOS-Alpha-One-Sentiment-Data.
10. http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators: chaikin_money_flow_cmf
11. Kuhn, M., Johnson, K.: Applied Predictive Modeling, Springer, 2013.
12. Malliaris, M.E., Salchenberger, L.: A neural network model for estimating option prices, Journal of Applied Intelligence 3(1993), 193–206.
13. Marechal, F., Stamate, D., Olaniyan, R, Marek, J.: On XLE Index Constituents' Social Media Based Sentiment Informing the Index Trend and Volatility Prediction, Proceedings of the 10th Intl. Conference on Computatational Collective Intelligence (ICCCI), Springer, LNCS, 2018.
14. McLeod, A., Xu, C.: bestglm: Best Subset GLM, https://CRAN.R-project.org/package=bestglm, 2010
15. Olaniyan,R., Stamate,D., Logofatu,D.: Social web-based anxiety index's predictive information on $S\&P$ 500 revisited, Proceedings of the 3rd Intl. Symposium on Statistical Learning and Data Sciences, Springer, LNCS, 2015.
16. Olaniyan,R., Stamate,D., Logofatu,D., Ouarbya,L.: Sentiment and Stock Market Volatility Predictive Modelling - a Hybrid Approach, Proceedings of the 2nd IEEE/ACM International Conference on Data Science and Advanced Analytics, 2015.
17. Rumelhart, D.E., McClelland, J.L.: Parallel distributed processing, MIT Press,Cambridge, MA, 1986.
18. Shiller, R.J.: Irrational Exuberance Princeton: Princeton University press, 2000.
19. Sprenger, T.O., Tumasjan,A., Sandner,P.G., and Welpe, I.M.: Tweets and trades: the information content of stock microblogs, European Financial Management, 20(5), 926–957, 2014.