Sentiment and stock market volatility predictive modelling - a hybrid approach

Rapheal Olaniyan, Daniel Stamate
and Lahcen Ouarbya
Data Science & Soft Computing Lab, and
Department of Computing
Goldsmiths College
University of London
Email:d.stamate@gold.ac.uk

Doina Logofatu
Department of Computer Science
Frankfurt University of Applied Sciences
Email:logofatu@fb2.fh-frankfurt.de

***Abstract*—The frequent ups and downs are characteristic to the stock market. The conventional standard models that assume that investors act rationally have not been able to capture the irregularities in the stock market patterns for years. As a result, behavioural finance is embraced to attempt to correct these model shortcomings by adding some factors to capture sentimental contagion which may be at play in determining the stock market. This paper assesses the predictive influence of sentiment on the stock market returns by using a non-parametric nonlinear approach that corrects specific limitations encountered in previous related work. In addition, the paper proposes a new approach to developing stock market volatility predictive models by incorporating a hybrid GARCH and artificial neural network framework, and proves the advantage of this framework over a GARCH only based framework. Our results reveal also that past volatility and positive sentiment appear to have strong predictive power over future volatility.**

***Index Terms*—Granger causality, non-parametric test, GARCH, EGARCH, artificial neural networks, sentiment, stock market, volatility, Monte Carlo simulations.**

I. Introduction

STANDARD finance models are built under the main assumption that investors act rationally. These models make use of conventional data like the stock market data. The models assume that stock market returns are equal to fundamental returns, where the market returns reflect all known information. In view of the assumptions of market efficiency and investor rationality, the Efficiency Market Hypothesis (EMH) became popular. This hypothesis adds substance to the traditional finance models as these reflect the idea that all new information has already been factored into the stock market prices.

After many years of model predictions, the models appear too basic judging by their inefficiency in capturing the complex and dynamic nature of the stock market: stock market returns and investor behaviour diverge away from the fundamental prices and rationality, respectively. These call for attention in behavioral finance to resolve the shortcomings of the standard finance models. Behavioral finance relaxes the assumption that investors act rationally. Since then, researchers have been focusing on the relationship between sentiment and the stock market. Shiller [20] and Sprenger et al. [24] opposed the EMH by stating 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.

Schumaker and Chen [22], Bollen et al. [13], Baker and Wurgler [15] and Gilbert and Kahahalios [9] investigated the causal relationship between the stock market returns and the sentiment and all reached the same conclusion that sentiment influences the stock market returns. Results from [9] based on a collection of LiveJournal blogs, showed that sentiment possesses predictive information on the stock market returns. However, the models these results were based upon, presented flaws from a statistical point of view, which were analysed and corrected by Olaniyan et al. in [21], which investigated the causality direction between sentiment and the stock market returns using a non-parametric approach. They showed that there is no line of Granger causality between the stock market returns and sentiment.

Undoubtedly, most of the research work attempting to uncover the relationship among stock market returns, volatility, sentiment, among others, do so with the aim that abnormal profits can be earned. What happens to the stock market volatility in the face of rising stock market prices and vice versa? Black [11] observed that a 1% summed return might result in more than 2% drop in volatility especially for a low volatility stock. Observation from the research work conforms with rationality in that investors consider risk to be positively related to volatility. The higher the negative sentiment, the higher the risk associated to a stock, the higher the stock volatility.

So far, the causality relationship between the stock market and sentiment has been investigated. But there is little evidence to support that sentiment resolves stock market uncertainty: as we will show here, evidence rather indicates that sentiment induces volatility. How can we predict the impacts of sentiment on stock market volatility? How can we investigate the asymmetric effects of different sentiments on the stock market volatility? Knowing that GARCH framework

978-1-4673-8273-1/15/$31.00 2015 IEEE

is popular in predicting the stock market volatility, how can we develop a much more efficient stock market predictive model by using the GARCH model as a benchmark? These are the main questions this paper focuses on.

Black [10] observed a negative correlation between current stock market returns and future return volatility because bad news tends to increase volatility as the realised return is lower than expected, and good news tends to reduce volatility as the realised return is higher than expected. Lee et al. [25] employed a generalized autoregressive conditional heteroscedasticity-in-mean specification to examine the impacts of investment sentiment on stock market returns and volatility. They emphasized that focusing alone on the impacts of sentiment either on the mean or variance in asset returns could lead to misspecification problems. A GARCH framework was used to analyse the effects, and results showed that shifts in sentiment are negatively correlated with market volatility. That is, volatility increases (decreases) when investors become more bearish (bullish).

In view of these areas of growing interest, our paper attempts to examine the relationship among stock market returns, volatility, and stock-related sentiment. Secondly, we investigate the asymmetric impacts of positive and negative stock-related sentiments on the stock market volatility. More so, we propose a much more efficient volatility predictive model that incorporates both an GARCH framework and an artificial neural network framework.

The remainder of this paper is organized as follows. Section 2 describes the non-parametric approach we use, and presents our results of the causality relationship between sentiment and the stock market returns. It also presents our benchmark volatility predictive model and assesses the asymmetric effects of positive and negative sentiments on the stock market volatility. Section 3 entails our new hybrid approach that incorporates both the GARCH framework and the artificial neural network framework. Section 4 reveals our findings and concludes the paper.

II. Stock market and sentiment

We use stationary daily time series variables obtained from stock market data and also stock-related sentiment to measure the influence sentiment has on the stock market returns. The S&P 500 index values from the 6th of September 2012 to 12th of May 2014 are used as a proxy for the stock market data, and are employed to generate two variables participating in the development of predictive models, namely the stock market acceleration metric denoted as M and the volume of stock trading denoted as V. 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 $M_t = R_{t+1} - R_t$. V_t is expressed as the first difference of the logged trading volume. The sentiment series S is obtained directly from StockTwits, which contains sentiment-filled S&P 500 blogs on Twitter (see the Downside Hedge website for more detailed explanation about the sentiment building process [6]). We now define $A_t = S_t - S_{t-1}$. Moreover, we include sentiment dummy variables so that we could measure the asymmetric impacts of positive and negative sentiments on the stock market volatility. We do not have access to these different sentiments. We resolve to using proxies for positive sentiment dummy variable $D_t = 1$ where $A_t - A_{t-1} > 0$ and 0 otherwise. We are able to generate the positive sentiment and negative sentiment series by defining $P_t = A_t^2 * D_t$ and $N_t = A_t^2 * (1 - D_t)$, respectively. Our volatility series Q is generated using the exponential GARCH(1,1), denoted also by EGARCH(1,1), as follows:

$$Q_t = ln(\sigma_t^2) = \omega + \beta ln(\sigma_{t-1}^2) + \alpha\Big[\frac{\mid \varepsilon_{t-1} \mid}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}}\Big] + \gamma\Big(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\Big) + \theta_1 P_{t-1} + \theta_2 N_{t-1}, \tag{1}$$

where β measures the impact of past volatility on future volatility, α measures the impact of positive stock market shock on the stock volatility, γ captures the impact of negative stock market shock on the stock volatility, and θ_1 and θ_2 measure the impacts of positive and negative sentiments, respectively, on the stock volatility.

A. Conventional Granger causality between sentiment and stock market returns

In the process of determining the causal relationship between sentiment and stock market returns we present the general linear VAR models as:

$$M1 : M_t = \alpha_1 + \Sigma_{i=1}^3 \beta_{1i} M_{t-i} + \Sigma_{i=1}^3 \gamma_{1i} V_{t-i} + \Sigma_{i=1}^3 \delta_{1i} Q_{t-i} + \epsilon_{1t} \tag{2}$$

$$M2 : M_t = \alpha_2 + \Sigma_{i=1}^3 \beta_{2i} M_{t-i} + \Sigma_{i=1}^3 \gamma_{2i} V_{t-i} + \Sigma_{i=1}^3 \delta_{2i} Q_{t-i} + \Sigma_{i=1}^3 \eta_{2i} A_{t-i} + \epsilon_{2t} \tag{3}$$

The models $M1$ and $M2$ are used to measure the influence of the sentiment on stock prices. The difference in the models is that $M1$ does not include the sentiment variable, it only uses the lagged market variables mentioned above in this section. $M2$ adds the lagged sentiment to the $M1$'s variables. If $M2$ performs better than $M1$, one could conclude that the sentiment has predictive information on the stock market. But such a conclusion is dependent on the conditions that the estimated residuals are normally distributed and homoscedastic in variance.

Before Eq. (2) and (3) can be estimated, the volatility series Q must be established and the influence of sentiment on volatility assessed. Does sentiment have predictive information on volatility? What asymmetric impacts do positive and negative stock market shocks have on volatility? What asymmetric impacts do positive and negative sentiments have on volatility? Solving Eq. (1) and (3) provides answers to the above questions.

The traditional volatility model is built using a GARCH approach that uses the residuals from a linear model as input to generate the volatility series.

TABLE I: Only parameters from Eq. (1) that are statistically significant are reported. $LjungBox_R$ and $LjungBox_{R^2}$ denote Ljung-Box tests on the standardised residuals and squared residuals respectively.

Variable	Estimate	t value	p-value
ω	-2.1849	-3.7214	< 0.001
β	0.7690	12.4479	< 0.001
λ	0.2926	4.2497	< 0.001
θ_1	-6.1031	-2.3453	0.0190
Test	$LjungBox_R$	$LjungBox_{R^2}$	$ARCHLM$
p-value	> 0.05	> 0.05	> 0.05

We start by introducing the general framework of our models. Consider a regression modeling with a constant conditional variance, $VAR(Y_t \mid X_{1,t}, ..., X_{m,t}) = \sigma_\epsilon^2$. Then regressing Y_t on $X_{1,t}, ..., X_{m,t}$ can be generally denoted as:

$$Y_t = f(X_{1,t}, ..., X_{m,t}) + \epsilon_t, \tag{4}$$

where ϵ_t is independent of $X_{1,t}, ..., X_{m,t}$ with expectation equal to 0 and constant conditional variance σ_ϵ^2. Here $f(\cdot)$ is the conditional expectation of $Y_t \mid X_{1,t}, ..., X_{m,t}$. Eq. (4) can be extended to include conditional heteroscedasticity as follows:

$$Y_t = f(X_{1,t}, ..., X_{m,t}) + \sigma(X_{1,t}, ..., X_{m,t})\epsilon_t, \tag{5}$$

where $\sigma^2(X_{1,t}, ..., X_{m,t})$ is the conditional variance of $Y_t \mid X_{1,t}, ..., X_{m,t}$ and ϵ_t has the mean 0 and the conditional variance 1. Since $\sigma(X_{1,t}, ..., X_{m,t})$ is a standard deviation, it is captured using a non-linear non-negative function in order to maintain its non-negative structure. This leads us to the traditional GARCH model defined as:

$$\sigma_t^2 = \omega_0 + \beta_1\sigma_{t-1}^2 + \alpha_1\epsilon_{t-1}^2 \tag{6}$$

The problem with Eq. (6) is that the asymmetric effects of different market shocks could not be captured. As a result, a new model was introduced by Nelson [3]. This model is called the Exponential GARCH model defined in Eq. (1) to capture these asymmetric effects of different shocks on the stock market volatility. This proposed model has earned popularity as it makes it possible to measure the asymmetric effects of market shocks. We use this model as our benchmark in predicting the stock market volatility.

In order to obtain the volatility series Eq. (3) is estimated without the variable Q and the model residuals are applied to Eq. (1) to generate Q. Table I presents the results of the estimated volatility model.

It is revealed that past volatility has positive relationship with regard to future volatility. In fact, it is observed that it influences future volatility the most. It is also shown that negative market shocks are positively related to market return volatility. They increase the level of market risk and therefore influence the stock volatility positively. The asymmetric impacts of different sentiments on stock volatility are also captured. As it would be expected, positive sentiment reduces volatility. Oddly, negative sentiment does not appear to be statistically important. Goodness of fit tests are also employed on the standardised residuals and squared residuals of the estimated EGARCH model. The insignificant p-values from the Ljung-Box tests on both the standardised residuals and squared residuals, and the ARCH LM test, suggest that the EGARCH model would fit the data well.

The volatility series obtained in Eq. (2) and (3) are estimated and the linear Granger causality test results are presented in Table II. The first two columns in the table show that $M2$, with the sentiment included in the analysis, would outperform $M1$, judging from the Granger causality F statistics $F_{3,401} = 6.5385$, and the corresponding p-value $p_{Granger} = 0.0003$.

TABLE II: Granger Causality results and Monte Carlo Simulation. $MCp_{Gausskern}$, and MCp_{boot} are the p-values of the simulations using a Gaussian kernel assumption, and bootstrap sampling respectively.

$F_{3,401}$	$p_{Granger}$	$MCp_{Gausskern}$	MCp_{boot}	Shapiro-Wilk
6.5385	0.0003	0.0005	0.0005	0.0047

However, there are some concerns in the estimated models: the estimated residuals possess serious autocorrelation, are non-normal and heteroscedastic in variance (having p-values Ljung-Box < 0.05 for lags equal to or greater than 3, and Shapiro-Wilk $= 0.0047$) and the heteroscedastic presence is revealed in the EGARCH process in Table I (with p-value of $\beta < 0.001$). These are major shortcomings of the linear Granger causality test results according to which sentiment would be a determining factor in predicting the stock market returns. In an attempt to see if we could still rely on the test results, Monte Carlo simulations with a Gaussian kernel distribution assumption for the sentiment series are employed. 1 million sets of samples are generated for the sentiment, and are fed into (3) by iterating 1 million times. The same number of F statistic values are generated in the process and then classified based on if the F statistic is at least 6.5385. The total number of F statistic values that are at least 6.5385 is then divided by the number of iterations to obtain the Monte Carlo experimental p-value $MCp_{Gausskern}$ = 0.0005 as shown in the third column of Table II.

Although $MCp_{Gausskern} = 0.0005$ seems to confirm the conclusion of the Granger causality analysis, the Monte Carlo simulation suffered at its turn of the issue of retrieving a significantly different experimental p-value with respect to $p_{Granger}$. This issue seems to be the consequence of another issue, consisting of the fact that the empirical distribution of the F-statistic computed in the Monte Carlo experiments significantly deviated from the expected F-distribution, as confirmed by the Kolmogorov-Smirnov test (D = 0.0348, $p < 0.001$).

This realization constitutes a nontrivial reason to question the Monte Carlo estimates, and a natural question which arises is: would the assumption of the Gaussian kernel distribution

for the sentiment have possibly introduced a bias in the simulation? To answer the question, we apply another non-parametric Monte Carlo simulation method based on the bootstrap sampling. We follow the same procedure as that used in the Gaussian Kernel Monte Carlo simulation. The result is presented in the fourth column of Table II, where MCp_{boot} denotes the p-value issued from the use of the bootstrap sampling method. The simulation led to a similar p-value of 0.0005. Also, the empirical distribution of the F-statistic computed in the bootstrap sampling Monte Carlo experiment is different from the expected F-distribution (Kolmogorov-Smirnov test result having D = 0.0351, $p < 0.001$). These shortcomings confirm once again that proving the relationship between the sentiment and stock market is problematic if linear models are involved.

Although there are strong reasons to accept the Granger causality results on one hand, there are also issues regarding the assumptions clearly stated under the linear regression modelling, such as the residuals must be independent, normally distributed, and homoscedastic in variance. All these assumptions are violated in our estimated models despite the fact that our Monte Carlo simulations fairly validate the Granger causality test results. As such, in the next sub-section we devise a non-parametric non-linear Granger causality test in the context of our problem, in an attempt to overcome the limitations illustrated in the present sub-section.

B. Extending the approach to non-parametric non-linear Granger causality

Stock market exhibits frequent volatility and this makes linear frameworks less capable in capturing and predicting its trends. For stock market predictive values to be considered reliable, two key necessary and sufficient requirements must be met. The first would be to generate an acceptable predictive model and the second would be to rigorously and statistically prove the model's predictive value. The inability of any model to satisfy these two conditions casts doubt on its predictive value. This has been the case with most research work attempting to examine the causality direction between the stock market and sentiment-filled online expressions. Gilbert and Karaholios are among the very few that attempted to statistically prove their models' predictive value in their highly cited work [9]. But their results appeared to be biased as a consequence of their non-normal estimated model residuals and heteroscedasticity. These results have finally been proved not to be valid by further investigation in subsequent work [21].

In this paper we apply a non-parametric statistical technique for detecting nonlinear causal relationships between the residuals of linear models, technique which was originally proposed by Baek and Brock [8] and was later modified by Hiemstra and Jones [2] to become one of the most popular techniques for detecting nonlinear causal relationships among variables.

Consider two series X_t and Y_t as follows: let the Lx and Ly be the lag length of the lag series X_{t-Lx}^{Lx} and Y_{t-Ly}^{Ly} of X_t and Y_t, respectively, and let us denote the k-length lead vector of Y_t by Y_t^k. In other words,

$$\begin{aligned} Y_t^k &\equiv (Y_t, Y_{t+1}, ..., Y_{t+k-1}), k \geq 1, t \geq 1 \\ Y_{t-Ly}^{Ly} &\equiv (Y_{t-Ly}, Y_{t-Ly+1}, ..., Y_{t-1}), Ly \geq 1, \\ &\qquad t = Ly+1, Ly+2, ... \\ X_{t-Lx}^{Lx} &\equiv (X_{t-Lx}, X_{t-Lx+1}, ..., Y_{t-1}), Ly \geq 1, \\ &\qquad t = Lx+1, Lx+2, ... \end{aligned} \tag{7}$$

Given arbitrary values for $k, Lx, Ly \geq 1$ and $\varepsilon > 0$, then X_t does not strictly nonlinearly Granger cause Y_t if:

$$\begin{aligned} Pr(\parallel Y_t^k - Y_s^k \parallel < \varepsilon \mid \parallel Y_{t-Ly}^{Ly} - Y_{s-Ly}^{Ly} \parallel < \varepsilon, \\ \parallel X_{t-Lx}^{Lx} - X_{s-Lx}^{Lx} \parallel < \varepsilon) = Pr(\parallel Y_t^k - Y_s^k \parallel < \\ \varepsilon \mid \parallel Y_{t-Ly}^{Ly} - Y_{s-Ly}^{Ly} \parallel < \varepsilon), \end{aligned} \tag{8}$$

where $Pr(A \mid B)$ denotes the probability of A given B, $\parallel \cdot \parallel$ is the maximum norm, i.e. for a vector $V \equiv (v_1, v_2, \ldots, v_m)$, $\parallel V \parallel = max\{|v_1|, \ldots, |v_m|\}$, $s, t = max(Lx, Ly)+1, \ldots, N-k+1$, N is the length of the time series and ε is N-dependent and typically has values between 0.5 and 1.5 after normalising the time series to unit variance. The left hand side in (8) is the conditional probability which implies that two arbitrary k-length lead vectors of Y_t are within a distance ε, given that two corresponding Lx- length lag vectors of X_t and two corresponding Ly-length lag vectors of Y_t are within a distance of ε. The right hand side in (8) is the probability that two arbitrary k-length lead vectors of Y_t are within a distance of ε, given that the two corresponding Ly-length lag vectors of Y are within the distance of ε.

Eq. (8) can be rewritten using conditional probabilities in terms of the ratios of joint probabilities as follows:

$$\frac{CI_1(k+Ly, Lx, \varepsilon)}{CI_2(Ly, Lx, \varepsilon)} = \frac{CI_3(k+Ly, \varepsilon)}{CI_4(Ly, \varepsilon)} \tag{9}$$

The joint probabilities are defined as:

$$\begin{aligned} CI_1(k+Ly, Lx, \varepsilon) &\equiv Pr(\parallel Y_{t-Ly}^{k+Ly} - Y_{s-Ly}^{k+Ly} \parallel < \varepsilon, \\ &\qquad \parallel X_{t-Lx}^{Lx} - X_{s-Lx}^{Lx} \parallel < \varepsilon) \\ CI_2(Ly, Lx, \varepsilon) &\equiv Pr(\parallel Y_{t-Ly}^{Ly} - Y_{s-Ly}^{Ly} \parallel < \varepsilon, \\ &\qquad \parallel X_{t-Lx}^{Lx} - X_{s-Lx}^{Lx} \parallel < \varepsilon) \\ CI_3(k+Ly, \varepsilon) &\equiv Pr(\parallel Y_{t-Ly}^{k+Ly} - Y_{s-Ly}^{k+Ly} \parallel < \varepsilon) \\ CI_4(Ly, \varepsilon) &\equiv Pr(\parallel Y_{t-Ly}^{Ly} - Y_{s-Ly}^{Ly} \parallel < \varepsilon) \end{aligned} \tag{10}$$

The Correlation-Integral estimators of the joint probabilities expressed in Eq. (10) measure the distance of realizations of a random variable at two different times. They are proportions defined as the number of observations within the distance ε to the total number of observations. Let us denote the time series of realizations of X and Y as x_t and y_t for $t = 1, 2, ..., N$ and let y_t^k, y_{t-Ly}^{Ly} and x_{t-Lx}^{Lx} denote the k-length lead, and Lx-length lag vectors of x_t and the Ly-length lag vectors of y_t as defined in Eq. (7). In addition, let $I(Z_1, Z_2, \varepsilon)$ denote a

TABLE III: Assigning values to ε, as of Diks and Panchenko [1]

n	100	200	500	1000	2000	5000	10,000	20,000	60,000
ε	1.5	1.5	1.5	1.2	1	0.76	0.62	0.51	0.37

TABLE IV: Non-linear Granger non-causality 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.

$Lx = Ly = 1$	$p - value$
$A => M$	0.66433
$M => A$	0.30186

kernel that equals 1 when two conformable vectors Z_1 and Z_2 are within the maximum-norm distance ε of each other, and 0 otherwise. The Correlation-Integral estimators of the joint probabilities in Eq. (10) can be expressed as:

$$\begin{aligned} CI_1(k+Ly, Lx, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_{t-Ly}^{k+Ly}, y_{s-Ly}^{k+Ly}, \\ &\quad \varepsilon) \cdot I(x_{t-Lx}^{Lx}, x_{s-Lx}^{Lx}, \varepsilon), \\ CI_2(Ly, Lx, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_{t-Ly}^{Ly}, y_{s-Ly}^{Ly}, \\ &\quad \varepsilon) \cdot I(x_{t-Lx}^{Lx}, x_{s-Lx}^{Lx}, \varepsilon), \\ CI_3(k+Ly, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_{t-Ly}^{k+Ly}, y_{s-Ly}^{k+Ly}, \varepsilon), \\ CI_4(Ly, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_{t-Ly}^{Ly}, y_{s-Ly}^{Ly}, \varepsilon), \end{aligned} \tag{11}$$

where $t, s = max(Lx, Ly) + 1, ..., N - k + 1$, $n = N + 1 - k - max(Lx, Ly)$.

Given that two series, X and Y, are strictly stationary and meet the required mixing conditions mentioned in Denker and Keller [16], under the null hypothesis that X does not strictly Granger cause Y, the test statistics T is asymptotically normally distributed and it follows that:

$$\begin{aligned} T = \sqrt{n}\Big(\frac{CI_1(k+Ly, Lx, \varepsilon, n)}{CI_2(Ly, Lx, \varepsilon, n)} - \frac{CI_3(k+Ly, \varepsilon, n)}{CI_4(Ly, \varepsilon, n)}\Big) \\ \sim N\big(0, \sigma^2(k, Ly, Lx, \varepsilon)\big), \end{aligned} \tag{12}$$

where $n = N+1-k-max(Lx, Ly)$ and $\sigma^2(\cdot)$, the asymptotic variance of the modified Baek and Brock test statistics, and an estimator for it are defined in the Appendix in Hiemstra and Jones [2].

To resolve the shortcomings of the linear Granger causality test, VAR models for stock market returns and sentiment are exploited. For stock market returns, we make use of (3) and for the sentiment-based model, we have:

$$\begin{aligned} A_t = c_3 + \Sigma_{i=1}^{3} h_{3i} M_{t-i} + \Sigma_{i=1}^{3} \gamma_{3i} V_{t-i} + \\ \Sigma_{i=1}^{3} \delta_{3i} Q_{t-i} + \Sigma_{i=1}^{3} \eta_{3i} A_{t-i} + \epsilon_{3t} \end{aligned} \tag{13}$$

Note that (3) and (13) are estimated and the residuals from the estimated models are applied to (12).

Diks and Panchenko [1] provided some important improvement to the non-linear Granger Causality test. They demonstrated that the value to be arbitrarily assigned to the distance ε is highly conditional on the length n of the time series. The larger the value n, the smaller the assigned value for ε and, the better and more accurate the results. Most of the related works choose $k = Lx = Ly = 1$. The length of the series we are analysing is less than 500, so choosing ε=1.5 conforms with Table III.

The results of the tests presented in Table IV show that sentiment does not have any predictive power on the stock market returns, as the corresponding p-value of 0.66433 doesn't show statistical significance. This is clearly contrary to the findings of the linear Granger causality tests which have been invalidated by the presence of residual non-normality and heteroscedasticity.

Having observed no causal relationship between sentiment and stock market returns, can one reach the same conclusion that sentiment has no predictive power over the stock market volatility? Is the EGARCH model used for volatility model efficient in reliably predicting the stock market volatility? We investigate these problems in the next section.

III. A HYBRID APPROACH TO PREDICTING STOCK VOLATILITY

In this section we will demonstrate the predictive power of sentiment on the stock market volatility by proposing a hybrid approach based on the GARCH framework and the artificial neural network framework in which we consider feed-forward and recurrent neural networks. However, in order to propose this hybrid approach, we start by simply attempting to assess the predictive influence of sentiment on the volatility using the EGARCH model alone first, and evaluate the relative improvements when we enhance our approach with feed-forward and Elman neural networks.

Monfared and Enke [23] recently proposed a hybrid approach that incorporated GJR GARCH and feed-forward neural networks (NNs) in predicting volatility. Their model was applied to conventional variables such as market return, and variance of ten NASDAQ indices. Their findings showed that incorporating NNs into the GARCH framework improves volatility predictive performance. But how accurate is the

GARCH framework employed in predicting the stock market? Can some nonconventional variables like sentiment improve the performance of predictive models? We answer these questions by presenting new models that combine both EGARCH and neural network (NN) models.

Advancement in information processing technology contributed to the birth of NNs. According to Malliaris and Salchenberger [19], NNs present the relationship between inputs and outputs using the architecture of human brain to process large information and detect patterns by interconnecting and organizing them in different layers for information processing purposes. These layers are formed by a set of processing elements or neurons. The layers are structured in hierarchy consisting of input layers, output layers, and hidden layers. The connections between nodes possess some weights defining the influence of the output from a node on the input to the node connected to the former. These weights are extracted from the training data employed in the process of learning the relationship between the inputs and the outputs. Each of the processing element is assigned with an activation level, specified by continuous or discrete values. For neurons in the input layers, their activation levels are determined from the response obtained in the input signals within the environment. For neurons in the hidden or output layers, their activation levels are defined as a function of the activation levels of the neurons connected to them and the corresponding weights. The functions are called transfer function which may be in form of a linear discriminant function with value 1 for a positive signal if the value of the function exceeds a threshold level and 0 otherwise. The function may also be continuously nondecreasing, as is the case of the sigmoid functions. A feed-forward NN, for example, has a one-directional signal flow of mapping the inputs into the outputs from the input layer to the output layer. The applications of NN family are very popular in areas such as classifications, predictions, and pattern recognition, among others.

The NN family have different parameters in their design and these parameters may alter their outputs. Therefore, they are designed for different research goals. The backpropagation is one of the most popular techniques to the areas of research work aforementioned. Collins et al. [7] applied it to underwriting problems. Malliaris and Salchenberger [17] also applied the backpropagation technique in estimating option prices. To determine the values for the parameters (weights in this case) in the algorithms, mean square error and gradient descent are employed. At each iteration, current parameters are updated by minimizing the mean square error based on the actual response values and desired response values. A detailed explanation of the process is provided by Rumelhart and McClelland [4]. Multilayer, feed-forward and recursive NNs, such as Jordan recursive NNs and Elman recursive NNs, have become popular and are sometimes preferred to the traditional feed-forward NNs in some types of applications, as for instance financial forecasting aplications.

We build our stock volatility predictive approach based on feed-forward and recurrent NNs combined with EGARCH model. In order to assess the impacts of sentiment on the performance of the prediction we consider two sets of input datasets. The first dataset contains only the lagged volatility series fitted by the EGARCH model, Q_{t-1}, and Q_{t-3}. The second dataset includes, in addition to the first dataset, lagged positive and negative sentiments $P_{t-1}, P_{t-2}, N_{t-1}$ and N_{t-2}. Q_{t-2} is excluded in both datasets because it is highly correlated with Q_{t-1} as presented in Table V.

We employ the series on various classes of NN models: the feed-forward NNs, the Elman recursive NNs and the Jordan recursive NNs. Knowing that the output of NN 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 optimised NN models have been generated, and the Root Mean Square Errors (RMSE) have been also obtained as presented in Tables VI and VII. The Elman recursive and the feed-forward NNs provide closely the same results and are clearly better than the Jordan recursive NNs. The RMSE from the feed-forward and Elman models in Table VII are lower than their corresponding RMSE in Table VI. This observation confirms the importance of including sentiment among the predictors of stock market volatility. The RMSE are clearly diminished when sentiment variables are included in the training. This observation is further investigated by employing the graphical representations of the trained NN models.

TABLE VI: Results from the use of dataset without sentiment series. $Size$ refers to the number of hidden units, Max denotes the number of iterations, $Weight$ denotes the weight decay and $RMSE$ is the root mean square error which is the square root of MSE.

NN	$Size$	Max	$Weight$	$RMSE$
Jordan	16	340		0.00017
Elman	24	1440		0.00010
feed-forward	29	1400	0.001	0.00011

TABLE VII: Results from the use of dataset with sentiment series. $Size$ refers to the number of hidden units, Max denotes the number of iterations, $Weight$ denotes the weight decay and $RMSE$ is the root mean square error which is the square root of MSE.

NN	$Size$	Max	$Weight$	$RMSE$
Jordan	20	1240		0.00017
Elman	30	1040		0.00004
feed-forward	30	1920	0.001	0.00005

Figure 1 presents the regression plots of our fitted volatility for the two datasets. The performance is judged by the closeness of the fitted volatility plot in red to the optimal line in black. The plots in the first column of Figure 1 represent charts from the dataset without sentiment variables, and the plots in the second column denote charts from the dataset

TABLE V: Correlation. Res denotes the response variable which is the present volatility. $Q1$, $Q2$, and $Q3$ are the past volatility variables representing Q_{t-1}, Q_{t-2} and Q_{t-3} respectively. $P1$, $P2$, $N1$ and $N2$ also represent variables P_{t-1}, P_{t-2}, N_{t-1} and N_{t-2} respectively.

	Res	$Q1$	$Q2$	$Q3$	$P1$	$P2$	$N1$	$N2$
Res	1.00000	0.34180	0.16608	0.07901	0.00636	-0.04535	0.02769	0.07474
$Q1$	0.34180	1.00000	0.66309	0.45433	-0.43650	-0.20726	0.11954	-0.06976
$Q2$	0.16608	0.66309	1.0000	0.66355	0.12456	-0.43668	-0.18678	0.11911
$Q3$	0.07901	0.45433	0.66355	1.00000	0.06874	0.12312	-0.09946	-0.18726
$P1$	0.00636	-0.43650	0.12456	0.06874	1.00000	-0.10964	-0.15795	0.28075
$P2$	-0.04535	-0.20726	-0.43668	0.12312	-0.10964	1.00000	0.39715	-0.15738
$N1$	0.02769	0.11954	-0.18678	-0.09946	-0.15795	0.39715	1.00000	-0.10480
$N2$	0.07474	-0.06976	0.11911	-0.18726	0.28075	-0.15738	-0.10480	1.00000

with sentiment variables included. The difference between the two datasets used are clearly presented by the feed-forward and the Elman NN models. The plots in the second column appear much better than those in the first column and this suggests that by including sentiment variables one produces better predictions. That is, sentiment plays an important role in predicting the stock market volatility.

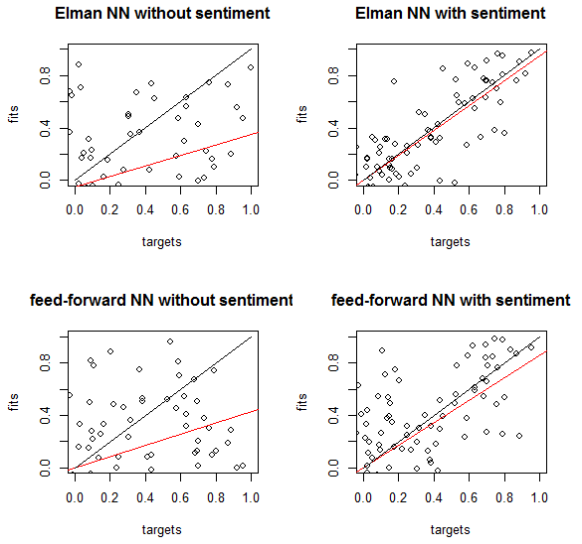

Fig. 1: Regression model. It presents information about the fitted volatility line in red and the optimal line in black, The plots in the first column represent plots of NN models with dataset without the sentiment variables. The plots in the second column represent volatility plots of dataset with sentiment variables.

Substantial evidence shows that sentiment has predictive information on the stock market. The previously produced EGARCH model has shown the significant importance of individual predictors. The relative importance of individual input variables in predicting stock market volatility is also investigated now in the context of our proposed hybrid approach combining EGARCH and NN models. The information contained in Figure 2 follows the same direction of interpretations as that presented in Table I. Past volatility

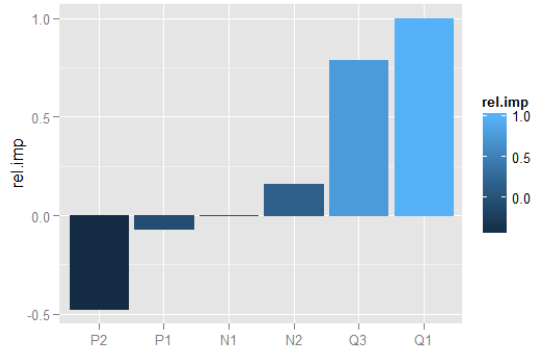

Fig. 2: Relative importance. It measures relative importance of the predictors in the model. Variables on the horizontal lines are the predictors. $Q1$, and $Q3$ are the past volatility variables representing Q_{t-1}, and Q_{t-3} respectively. $P1$, $P2$, $N1$ and $N2$ also represent variables P_{t-1}, P_{t-2}, N_{t-1} and N_{t-2} respectively. The variables with values below 0 have negative relationship with the response variable and those with values above 0 have positive relationship with the response variable. The response variable is the future volatility.

influences future volatility the most and it is positively related to future volatility. Positive sentiment has negative relationship with future volatility. Of all the predictors, negative sentiment appears to have the least influence on the stock volatility.

The visual interactions among the input, output and hidden nodes in the feed-forward NN model we built are also presented in Figure 3. In particular the figure illustrates if the relationship between the connected nodes are negative, positive, strong and weak. We notice the difference in the line thickness which expresses the magnitude of the weights.

A. *Sensitivity analysis*

We have shown how individual variables impact the response variable. Our findings present sentiment variables to be influential in predicting the stock market volatility. We have also shown the direction of the influence each variable has in predicting the stock market volatility. Recalling from the benchmark GARCH model used, past volatility has positive impact on future volatility. Positive sentiment has negative

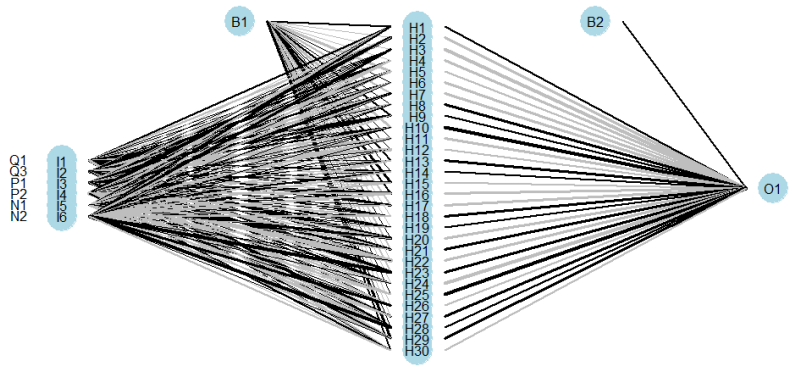

Fig. 3: Feed-forward NN nodes interaction plot. It depicts the interrelationship among the input, output and the hidden layers. The black lines represent positive weights, the grey lines represent negative weights, and the line thickness represents the magnitude of the weights on the connections.

influence on the stock market volatility. From the relative importance information of the explanatory variables presented in Figure 2 it is observed that positive sentiment and past volatility have higher impacts on the volatility just the same way as presented from the results obtained in the GARCH model. Some of the most relevant questions about the relationship between sentiment and the stock market variables have been answered from a rigorous / statistical point of view. Equally important is the proposed hybrid approach that derives a larger efficiency from the combination of GARCH framework and neural network framework in developing more advanced volatility predictive models. We have shown that our proposed model is highly efficient to this regard. Yet, some important questions are still left unanswered.

For a basic linear regression model, it is easy to observe how each explanatory variable impacts the response variable by keeping the other explanatory variables constant. Secondly, it provides categorical information about the direction of relationships between individual explanatory variables and the response variable. Being categorical implies that a relationship shows if an explanatory variable has positive or negative influence on the response variable, and this direction of relationship is constant. Clearly, linear models are simple, categorical and straight-forward. This brings forward the question: is there any way to present the form of relationship of every individual explanatory variable with the response variable from the proposed hybrid approach? That is, given the other explanatory variables constant, what amount of change will be impacted on the response variable for a unit change in an explanatory variable?

Neural networks are considered as a 'black box' as they do not offer insightful explanation about the impacts of individual input variables in the prediction process. Gevrey et al. [18] are among the early researchers that provided these long-awaited insights. They make it possible to carry out sensitivity analysis on these individual explanatory variables. In order for us to answer the pressing question about the form of relationship of each explanatory variable on the response variable we use techniques of sensitivity analysis. In particular, we intend to examine how a unit change in each explanatory variable influences the response variable. We also aim to examine if the relationship between an explanatory variable and a response variable changes with regard to the constant values of all the other explanatory variables.

Figure 4 presents our sensitivity analysis results. Each of the 6 columns corresponds to each explanatory variable whose label is provided at the top of each column. At the far-right of Figure 4 we have $Splits$, which denotes the different constant values assigned to the other explanatory variables while one explanatory variable is under consideration with respect to the response variable. For each constant value there is a corresponding colour line as illustrated in the figure.

First, we determine the relative importance of the individual explanatory variables based on the slope of the curves. Starting with the first column with label $N1$, when the other variables are kept constant at value 0, the top most line denotes the relationship of the negative sentiment variable $N1$ with respect to the response variable. It is observed that there is a negative relationship between these two variables, up until $N1$ is 0.5. At this stage, the relationship is inelastic. That is, less than a unit change is expected in the sentiment variable $N1$ to cause a unit change in the response variable. From the point where $N1$ is 0.5 and above the form of the relationship changes to positive, and the relationship is elastic. This shows that the form of relationship between the explanatory variable and the response variable may not necessarily be constant over time.

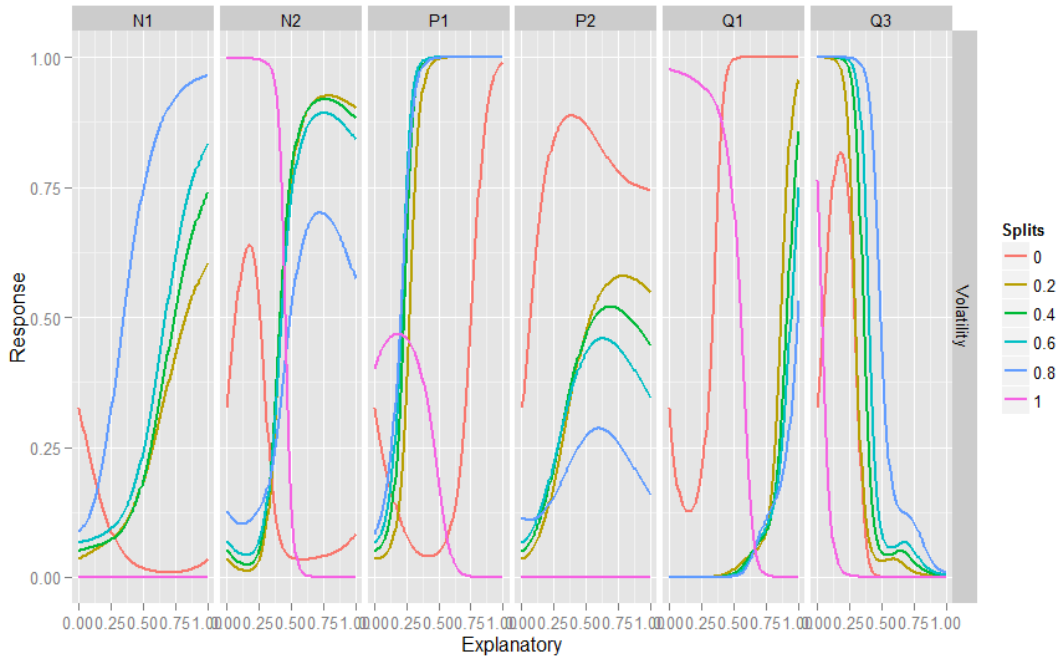

Fig. 4: Sensitivity analysis plots. It depicts the forms of relationship between each explanatory variable with regard to the response variable while keeping other explanatory variables constant. $N1$ and $N2$ denote first and second lagged negative sentiment variables respectively. $P1$ and $P2$ denote first and second lagged positive sentiment variables respectively. $Q1$ and $Q2$ are first and third lagged volatility variables. The dataset are normalised to be between 0 and 1.

In the second column corresponding to the negative sentiment variable $N2$, we observe a positive relationship up to the point where the value of $N2$ is around 0.2 when the other explanatory variables are held constant at value 0. When it has values between 0.25 and 0.5, the direction of relationship changes to negative and it is also inelastic which means that less than a unit change in $N2$ is expected to cause a unit change in the response variable. Above the value 0.5, the relationship changes again. This also confirms that the form of relationship using our hybrid approach is not constant and that may be the case with most neural network models. Comparing the two negative sentiment variables $N1$ and $N2$, all lines of $N2$ except the blue line are longer than those of $N1$, which means that the relationships for $N2$ are generally more inelastic (less elastic) than those for $N1$. In terms of these variables' relative importance, $N2$ is therefore more important than $N1$. Interestingly, it is revealed that the form of relationship between an explanatory variable with respect to the response variable differs for different constant values for all the other explanatory variables. When the other explanatory variables are kept constant at the maximum value 1, column 1 shows a change in relationship and it reveals that $N1$ does not have any influence on the response variable. At this stage, variables $N2$ and $Q1$ appear to be the most important predictors in relation to other predictors. When the constant values are set between 0.2 and 0.8, $P1$ and $Q3$ are the most important predictors with very strong inelastic relationships. Comparing all the explanatory variables relatively, $P1, Q1$ and $Q3$ are the most influential predictors. Recalling from Figure 2 and the EGARCH results in Table I that show positive sentiment to be negatively related with volatility, the information presented in Figure 4 does not disprove this form of relationship. These results from Figure 2 and the estimated EGARCH model are retrievable under some constant values of other explanatory variables. This analysis underscores the importance of sentiment variables in developing stock market volatility predictive models. Considering stock-related online expressions, our sensitivity analysis supports the finding that positive sentiment is significantly influential in predicting the stock market volatility. We have shown also that individual explanatory variables do not necessarily have a constant form of relationship with respect to the response variable. Different values of the other explanatory variables may cause an explanatory variable to change the form of relationship it has with the response variable.

IV. Discussion and Conclusion

This paper proposes a new approach to developing stock market volatility predictive models by incorporating a hybrid GARCH and artificial neural network framework. It also details the relationship among sentiment, stock market returns and volatility by applying a non-parametric non-linear Granger causality framework to assess the causality direction between sentiment and the stock market returns.

Our estimated EGARCH model shows that future volatility is influenced by factors such as past volatility and positive

sentiment. Negative sentiment from stock-related news does not appear to have any influence on volatility. The RMSE obtained from the EGARCH model is 9.724997. This is used as a benchmark to compare the efficiency of our proposed hydrid model. The RMSE is reduced to 0.0005 by our proposed model. It clearly confirms the superiority of our model over the benchmark. The model also reveals the relative importance and directional influence of individual variables.

We are able to show the asymmetric impacts of positive and negative sentiments on volatility using both the conventional and our hybrid models. Our results show that positive sentiment has statistical importance on the volatility and the relationship is negative. This finding goes in line with the conclusion from Black [10], Black [11] and Lee et al. [25]. When investors are optimistic (bullish) about a stock, the stock price increases and the volatility decreases because investors perceive the stock to have lower risk. Also, stock price would be expected to drop as a result of bad news which increases the associated stock risk and therefore results in increased volatility. But from our results, negative sentiment does not seem to influence volatility. This implies that the state of the economy may also be looked at in providing more insight into the relationship between sentiment and the stock market. For example, Olaniyan et al. [21] used stock market data during global recession and their findings proved negative sentiment to be statistically significant.

We employ sensitivity analysis to examine the form of relationship each explanatory variable has with respect to the stock market volatility from our hybrid model. Our results show that the form of relation could be positive, negative, bi-modal or come in any kind of form. It all depends on the values of the other explanatory variables employed at a point in time. Regardless, our sensitivity analysis shows that positive sentiment possesses predictive power on the stock market volatility. It also shows that past volatility impacts future volatility.

In conclusion, we have shown that sentiment built-up process is a determining factor when measuring the effects of sentiment on stock market volatility. [21] used sentiments that are not stock related and showed that past volatility and negative sentiment influence the stock market volatility. [21] observed also that positive sentiment does not have any significant impact on the volatility. But with our sentiments generated from stock-related blogs, past volatility still appears to have the strongest effect on future volatility. Positive sentiment has more effects on the stock market volatility than the negative sentiment. In fact, negative sentiment does not possess any significant statistical power on the stock market volatility. This implies that the sources of sentiments used also may have importance and therefore one must pay attention to the sources of sentiments used in developing stock market predictive models.

REFERENCES

[1] C. Diks, and V. Panchenko, *A new statistic and practical guidelines for nonparametric Granger causality testing*, Journal of Economic Dynamics and Control, 30(9-10), 1647–1669, 2006.

[2] C. Hiemstra, J.D. Jones, *Testing for linear and nonlinear Granger causality in the stock price–volume relation*, Journal of Finance, 49, 1639-1664, 1994.

[3] D.B. Nelson, *Conditional heteroscedasticity in asset returns: a new approach*, Econometrica, Vol.59, 347–370, 1991.

[4] D.E. Rumelhart, and J.L. McClelland, *Parallel distributed processing*, MIT Press,Cambridge, MA, 1986.

[5] D.M.R. Jason, S. Lawrence, T. Jaime, and R.K. David, *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*, In Proceedings of the Twentieth International Conference on Machine Learning, 616–623, 2003.

[6] Downside Hedge, *Twitter indicator for stock market analysis* www.downsidehedge.com/twitter-indicators/

[7] E. Collins, S. Ghosh, and C. Scofield, *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.

[8] E. Baek and W. Brock, *A general test for nonlinear Granger causality: bivariate model*, Working paper. Iowa State University, 1992.

[9] E. Gilbert, and K. Karahalios, *Widespread worry and the stock market*, In Proceedings of the 4th International Conference on Weblogs and Social Media, 58–65, 2010.

[10] F. Black, *The price of commodity contracts*, Journal of Financial Economics, 3, 267–179

[11] F. Black, *Studies of stock market volatility changes*, Proceedings of the American Statistical Association, Business and Economic Statistics Section, 177–181, 1976

[12] F. Yoav, and E.S. Robert, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Science, 49, 119–139 1997.

[13] J. Bollen, H. Mao, and X. Zeng, *Twitter mood predicts the stock market*, Journal of Computational Science, 2(1), 1–8, 2011.

[14] K. Balog, M. Gilad, R. de Maarten, *Why are they excited? Identifying and explaining spikes in blog mood levels*, In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[15] M. Baker and J. Wurgler, *Investor sentiment in the stock market*, Journal of Economics Perspectives, 21(2), 129–151.

[16] M. Denker, and G. Keller, *On U-statistics and von-Mises statistics for weakly dependent processes*, Z.Wahrscheinlichkeitsthorie und Verwandte Gebiete, 64, 505–522, 1983

[17] M.E. Malliaris, and L. Salchenberger, *A neural network model for estimating option prices*, Journal of Applied Intelligence 3(1993), 193–206.

[18] M. Gevrey, I. Dimopoulos, S. Lek, *Review and comparison of methods to study the contribution of variables in artificial neural network models*, Ecol. Model, 160: 249-264,2003.

[19] M. Malliaris, and L. Salchenberger, *Using neural networks to forecast the S$P 500 implied volatility*,Neurocomputing 10, 183–195, 1996.

[20] R.J. Shiller, *Irrational Exuberance* Princeton: Princeton University press, 2000.

[21] R. Olaniyan, D. Stamate, and D. Logofatu, *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, 2015.

[22] R.P. Schumaker, and H. Chen, *Textual analysis of stock market prediction using breaking financial news: The AZFin text system*, ACM Transactions on Information Systems, 27(2), 12:1–19, 2009.

[23] S.A. Monfared, and D. Enke, *Volatility forecasting using a hybrid GJR-GARCH neural network model*, Conference organized by Missouri University of Science and Technology, 2014–Philadelphia.

[24] T.O. Sprenger, A. Tumasjan, P.G. Sandner, and I.M. Welpe, *Tweets and trades: the information content of stock microblogs*, European Financial Management, 20(5), 926–957, 2014.

[25] W.Y. Lee, C.X. Jiang, and D.C. Indro, *Stock market volatility, excess returns, and the role of investment sentiment*, Journal of Banking and Finance, 26, 2277–2299, 2002