Social Web-based Anxiety Index's Predictive Information on S&P 500 Revisited

Rapheal Olaniyan[1], Daniel Stamate[1] and Doina Logofatu[2]

[1] Data Science & Soft Computing Lab, Department of Computing,
Goldsmiths College, University of London
[2] Department of Computer Science,
Frankfurt University of Applied Sciences

Abstract. There has been an increasing interest recently in examining the possible relationships between emotions expressed online and stock markets. Most of the previous studies claiming that emotions have predictive influence on the stock market do so by developing various machine learning predictive models, but do not validate their claims rigorously by analysing the statistical significance of their findings. In turn, the few works that attempt to statistically validate such claims suffer from important limitations of their statistical approaches. In particular, stock market data exhibit erratic volatility, and this time-varying volatility makes any possible relationship between these variables non-linear, which tends to statistically invalidate linear based approaches. Our work tackles this kind of limitations, and extends linear frameworks by proposing a new, non-linear statistical approach that accounts for non-linearity and heteroscedasticity.

1 Introduction

According to the investment theory, stock market is operating under the Efficient Market Hypothesis (EMH), in which stock prices are assumed to incorporate and reflect all known information. Sprenger et al. [15] strongly disagree with EMH by saying that the market is inefficient and therefore abnormal returns can be earned. In search for abnormal earning, researchers now 'listen' to news and mine online aggregated social data all in the course for these attractive profits.

Schumaker and Chen [14] are among the early researchers to investigate whether emotions can predict the stock market. Machine learning algorithms such as SVM, Naive Bayes, etc, are utilised to develop predictive models used to claim that financial news have a statistically significant impact on the stock market. Bollen et al. [4] present an interesting machine learning based approach to examine if emotions influence stock prices. Their results support the claim that emotions do influence the stock market.

The linear Granger causality analysis is employed by Gilbert and Karahalios [8] as a method to illustrate that web blog contained sentiment has predictive information on the stock market, but this method proved to have clear limitations as explained later in this paper. A linear model and the Granger causality

test are used also by Mao et al. [12] to examine the influence of social blogs on the stock market. The authors do raise some concerns about the possible non-linear nature in the relationship, but such concerns are not further explored. The non-linear Granger causality test, which relies on a Self-Organising Fuzzy Neural Network model, is unpopular in this area of work as it is thought not to be strong enough to capture volatile stock market movements, as revealed by Jahidul et al. [10]. Mittal and Goel [13] use machine learning algorithms to investigate if stock blogs, as a proxy for news, can predict this complex financial movement. Their findings make the same claim that stock blogs can be used to predict stock prices, and they use some level of accuracy of the predictive models to support their results.

Stock market is highly volatile. Therefore, capturing its movement and identifying relationships between stock prices and possible predictive variables require the use of appropriate approaches. These approaches should normally meet two requirements. The first requirement is to generate models for prediction, and the second requirement is to rigorously prove the models' predictive value.

As illustrated earlier in this section, there is a growing research work trying to establish that online expressed emotions have predictive information on the stock market. Most of these works fulfill the first requirement by devising and proposing various predictive models, but very few works attempt to fulfill also the second requirement by rigorously / statistically proving the predictive value of these models. Gilbert and Karahalios [8] are among the very few that do consider both requirements, by proposing a statistical approach, which is based on the Granger causality analysis and Monte Carlo simulations. We recognise the large interest and potential generated by [8] in inspiring further research that demonstrates the link between the online expressed emotions and the stock market. Our work builds upon the approach presented in [8], and does so by critically analysing it, by clearly identifying its drawbacks and limitations, by tackling these limitations and by extending the approach and the results presented in the paper. As such, we establish our findings on data which has been obtained from the [8] 's authors website.

The remainder of this paper is organized as follows. Section 2 briefly revisits the empirical analysis of Gilbert and Karahalios [8]. In particular it presents the data, and the Anxiety Index's building process. In addition, we discuss the essential limitations of the approach of [8], and provide and discuss the results of our alternative Monte Carlo simulations. Section 3 presents our new statistical based approach which captures efficiently the stock market volatility, and the predictive information relationship direction between stock prices and emotion. Section 4 entails our findings and conclusion.

2 Discussion on the Web blog based Anxiety Index

Four stationary daily time series variables were explored in Gilbert and Karahalios [8]: the Anxiety Index (AI), the stock return, and two control variables which are the trading volume and the stock volatility. All the variables were

generated from the stock market data S&P 500, except for the Anxiety Index AI.

[8] introduced the Anxiety Index using 20 million posts and blogs from LiveJournal, that had been gathered within three periods of 2008: January 25th to June 13th; August 1st to September 30th, and November 3rd to December 18th. Two sets of linguistic classifiers trained with a LiveJournal mood corpus from 2004 were employed to build the Anxiety Index metric. First, a corpus of 624,905 mood-annotated LiveJournal posts from Balog et al. [3] was used. 12,923 posts that users tagged as 'anxious', 'worried', 'nervous' or 'fearful' were extracted. Then two classifiers were trained to distinguish between 'anxious' and 'non anxious' posts. The first classifier $C1$, which was a boosted decision tree, as introduced by Yoav and Robert [16], used the most informative 100 word stems as features. The second classifier $C2$ consisting of a bagged Complement Naive Bayes model [11], used 46,438 words obtained from the 2004 corpus mentioned above. $C1_t$ and $C2_t$ were defined as the standard proportions of posts classified as 'anxious' by $C1$ and $C2$, respectively, during the closing trading day t. $C1_t$ and $C2_t$ were integrated in the series C defined by $C_t = max(C1_t, C2_t)$. The Anxiety Index was finally defined as the series $A_t = log(C_{t+1}) - log(C_t)$. 174 values were generated for this series from the available data.

The S&P 500 index was used as a proxy for the stock market, and was employed to generate three variables participating in the development of predictive models, namely the stock market acceleration metric denoted as M, the return volatility denoted as V, and the volume of stock trading denoted as Q. The stock return at time t was defined as $R_t = log(SP_{t+1}) - log(SP_t)$, where SP is the closing stock price. The stock market acceleration metric was obtained from the stock return as $M_t = R_{t+1} - R_t$. The stock return volatility was expressed as $V_t = R_{t+1} * R_{t+1} - R_t * R_t$, and finally Q_t was expressed as the first difference of the lagged trading volume.

2.1 Findings and limitations

The two OLS models employed by Gilbert and Karahalios in [8] are:

$$M1 : M_t = \alpha + \Sigma_{i=1}^{3}\beta_i M_{t-i} + \Sigma_{i=1}^{3}\gamma_i V_{t-i} + \Sigma_{i=1}^{3}\delta_i Q_{t-i} + \epsilon_t \quad (1)$$

$$M2 : M_t = \alpha + \Sigma_{i=1}^{3}\beta_i M_{t-i} + \Sigma_{i=1}^{3}\gamma_i V_{t-i} + \Sigma_{i=1}^{3}\delta_i Q_{t-i} + \Sigma_{i=1}^{3}\eta_i A_{t-i} + \epsilon_t \quad (2)$$

The models $M1$ and $M2$ were used to measure the influence of the Anxiety Index on stock prices. The difference in the models is that $M1$ does not include the Anxiety Index variable, it only uses the lagged market variables mentioned above in this section. $M2$ adds the lagged Anxiety Index to the $M1$'s variables. If $M2$ performs better than $M1$, one could conclude that the Anxiety Index has predictive information on the stock market. The first two columns of Table 1 show that $M2$, with the Anxiety Index included in the analysis, would outperform $M1$, judging from the Granger causality F statistics $F_{3,158} = 3.006$, and the corresponding p-value $p_{Granger} = 0.0322$.

Table 1. Granger Causality results and Monte Carlo Simulation. $MCp_{Gausskern}$, MCp_{inv} and MCp_{boot} are the p-values of the simulations using a Gaussian kernel assumption, the inverse transform sampling, and bootstrap sampling respectively.

$F_{3,158}$	$p_{Granger}$	$MCp_{Gausskern}$	MCp_{inv}	MCp_{boot}
3.006	0.0322	0.045	0.045	0.045

The main disadvantage of the approach of Gilbert and Karahalios [8] was that the Granger causality analysis's linear models M1 and M2 were actually not valid from a statistical point of view. In particular these models suffered of major shortcomings as for instance residuals were non-normally distributed, and they presented a heterogeneity of the variance. As such, although the p-value $p_{Granger} < 0.05$ suggests that the Anxiety Index adds significantly some predictive information on the stock market, such a conclusion is not supported by a valid statistical reasoning.

Due to the mentioned pitfalls, [8] proposed also a Monte Carlo simulation with a Gaussian kernel distribution assumption for the Anxiety Index, in an attempt to retrieve the same conclusion as in the non-statistically supported Granger causality analysis. The authors generated 1 million sets of samples for the Anxiety Index. These new series were used in (2) by iterating 1 million times to generate the same number of F statistic values, and then to classify these values based on if any F statistic is at least 3.01. The total number of F statistic's values that were at least 3.01 was then divided by the number of iteration to obtain the Monte Carlo experimental p-value, $MCp_{Gausskern} = 0.045$, shown in Table 1.

Although $MCp_{Gausskern} < 0.05$ seemed 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 seemed 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, i.e. $D = 0.0337,\ p < 0.001$ [8].

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 Anxiety Index have possibly introduced a bias in the simulation? To answer the question, we apply other non-parametric Monte Carlo simulation methods based on the inverse transform sampling method using the continuous version of the empirical distribution function corresponding to the original Anxiety Index's sample, and bootstrap sampling. We follow the same procedure as that used in [8]. Our Monte Carlo p-values are presented in the columns four and five of Table 1, where MCp_{inv} and MCp_{boot} denote p-values issued from the use of the inverse transform sampling and the bootstrap sampling methods. Both simulations led to a similar value of 0.045. Moreover, in both cases the empirical distribution of the F-statistic computed

in the Monte Carlo experiments is different from the expected F-distribution. These shortcomings confirm once again that proving the relationship between the Anxiety Index and stock prices is problematic if linear models are involved.

To this end we propose a new statistical approach to solve the limitations in [8] and to also reveal the relationship direction between the variables of interest.

3 Anxiety Index's predictive information on the stock market, revisited

We follow the guidelines from Diks and Panchenko [6] (see [7] for detailed explanation and software) to examine the line of Granger causality between the variables involved in our analysis. The idea of the non-parametric statistical technique for detecting nonlinear causal relationships between the residuals of linear models was proposed by Baek and Brock [2]. It was later modified by Hiemstra and Jones [9] and this has become one of the most popular techniques for detecting nonlinear causal relationships in 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} and Y_t^{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,

$$\begin{aligned} Y_t^k &\equiv (Y_t, Y_{t+1}, ..., Y_{t+k-1}), k = 1, 2, ..., t = 1, 2, .., \\ Y_t^{Ly} &\equiv (Y_{t-Ly}, Y_{t-Ly+1}, ..., Y_{t-1}), Ly = 1, 2, ..., t = Ly + 1, Ly + 2, ..., \\ X_t^{Lx} &\equiv (X_{t-Lx}, X_{t-Lx+1}, ..., Y_{t-1}), Ly = 1, 2, ..., t = Lx + 1, Lx + 2, ..., \end{aligned} \tag{3}$$

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

$$\begin{aligned} Pr(\parallel Y_t^k - Y_s^k \parallel < \varepsilon \mid \parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon) \\ = Pr(\parallel Y_t^k - Y_s^k \parallel < \varepsilon \mid \parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon) \end{aligned} \tag{4}$$

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 (4) is the conditional probability which implies that two arbitrary k-length lead vectors of Y_t are within a distance ε, given that two associating Lx- length lag vector of X_t and two associating Ly-length lag vector of Y_t are within a distance of ε. The right hand side in (4) 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 vector of Y are within the distance of ε.

Eq.(4) can be rewritten using conditional probabilities in terms of the ratios of joint probabilities as follows:

$$\frac{CI(k+Ly,Lx,\varepsilon)}{CI(Ly,Lx,\varepsilon)} = \frac{CI(k+Ly,\varepsilon)}{CI(Ly,\varepsilon)} \tag{5}$$

The joint probabilities are defined as:

$$\begin{aligned} CI(k+Ly,Lx,\varepsilon) &\equiv Pr(\parallel Y_t^{k+Ly} - Y_s^{k+Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon), \\ CI(Ly,Lx,\varepsilon) &\equiv Pr(\parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon), \\ CI(k+Ly,\varepsilon) &\equiv Pr(\parallel Y_t^{k+Ly} - Y_s^{k+Ly} \parallel < \varepsilon), \\ CI(Ly,\varepsilon) &\equiv Pr(\parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon) \end{aligned} \tag{6}$$

The Correlation-Integral estimators of the joint probabilities expressed in Eq. (6) 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} and x_t^{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 (3). In addition, let $I(Z_1, Z_2, \varepsilon)$ denote a 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 equation (6) can be expressed as:

$$\begin{aligned} CI(k+Ly,Lx,\varepsilon,n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{k+Ly}, y_s^{k+Ly}, \varepsilon) \cdot I(x_t^{Lx}, x_s^{Lx}, \varepsilon), \\ CI(Ly,Lx,\varepsilon,n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{Ly}, y_s^{Ly}, \varepsilon) \cdot I(x_t^{Lx}, x_s^{Lx}, \varepsilon), \\ CI(k+Ly,\varepsilon,n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{k+Ly}, y_s^{k+Ly}, \varepsilon), \\ CI(Ly,\varepsilon,n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{Ly}, y_s^{Ly}, \varepsilon), \end{aligned} \tag{7}$$

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 [5], under the null hypothesis that X does not strictly Granger cause Y, the test statistics T is asymptotically normally distributed and it follows that:

$$T = \sqrt{n}\Big(\frac{CI(k+Ly,Lx,\varepsilon,n)}{CI(Ly,Lx,\varepsilon,n)} - \frac{CI(k+Ly,\varepsilon,n)}{CI(Ly,\varepsilon,n)}\Big) \sim N\big(0, \sigma^2(k,Ly,Lx,\varepsilon)\big) \tag{8}$$

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 [9].

To test our variables for a possibly non-linear relation, 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:

$$Y_t = f(X_{1,t}, ..., X_{m,t}) + \epsilon_t, \tag{9}$$

where ϵ_t is independent of $X_{1,t}, ..., X_{m,t}$ with expectation zero and constant conditional variance σ_ϵ^2. $f(\cdot)$ is the conditional expectation of $Y_t \mid X_{1,t}, ..., X_{m,t}$. Eq.(9) can be extended to include conditional heteroscedasticity as follows:

$$Y_t = f(X_{1,t}, ..., X_{m,t}) + \sigma(X_{1,t}, ..., X_{m,t})\epsilon_t \tag{10}$$

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 GARCH models. Comparing Eq.(9) and Eq.(10), the first part of the right hand side of Eq.(9) is the same with that of Eq.(10). This is a linear model. The second part of the right hand side of Eq.(9) are residuals of the linear process. They represent the second part of the right hand side of Eq.(10). Eq.(9) can finally be presented in the VAR framework as:

$$\begin{aligned} M_t = c + \Sigma_{i=1}^3 h_i M_{t-i} + \Sigma_{i=1}^3 \gamma_i V_{t-i} + \Sigma_{i=1}^3 \delta_i Q_{t-i} + \\ \Sigma_{i=1}^3 \eta_i A_{t-i} + a_t \end{aligned} \tag{11}$$

$$\begin{aligned} A_t = c + \Sigma_{i=1}^3 h_i M_{t-i} + \Sigma_{i=1}^3 \gamma_i V_{t-i} + \Sigma_{i=1}^3 \delta_i Q_{t-i} + \\ \Sigma_{i=1}^3 \eta_i A_{t-i} + a_t \end{aligned} \tag{12}$$

Following the second part of the right hand side of Eq.(10), the residuals a_t from Eq.(11) and Eq.(12) are presented in GARCH(1,1) as:

$$a_t = \sigma_t \epsilon_t \tag{13}$$

where $\sigma_t = \sqrt{w + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2}$, in which w, α_1 and β_1 are constants. We finally derive the GARCH(1,1)-filtered residuals, standardized residuals, as

$$\epsilon_t = \frac{a_t}{\sigma_t} \tag{14}$$

We obtain the residuals from the VAR model in Eq. (11) and (12). The test statistic in Eq. (8) is then applied to these residuals to detect the causal relation between the Anxiety Index and stock prices. Diks and Panchenko [6] provide some important improvement to the Non-linear Granger Causality test. [6] demonstrates 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.

Table 2. Assigning values to ε, as of Diks and Panchenko [6]

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

Most of the related works choose $k = Lx = Ly = 1$. The length of the series we are analysing is less than 200, so choosing ε=1.5 conforms with Table 2. Given $\varepsilon = 1.5$, $k = Lx = Ly = 1$, the results from the test are presented in Table 3.

Our first result in this framework seems to support the idea that the Anxiety Index has predictive information on the stock market, as this is based on the p-value of 0.017 shown in the first row of Table 3. Some re-considerations are necessary though.

Hiemstra and Jones [9] state that the non-linear structure of series is related to ARCH errors. Anderson [1] proves that the volatility of time series contains predictive information flow. But Diks and Panchenko [6] warn that the presence of conditional heteroscedasticity in series could produce spurious results. To avoid any possible bias in our results, the residuals are applied to Eq.(13) to filter out any conditional heteroscedasticity in the residuals of the VAR models. We also rely on the GARCH(1,1)-filtered residuals to re-establish our findings.

We are able to identify, using the GARCH(1,1) results, that a_t from Eq.(11) is a GARCH process with ϵ_t being a Gaussian white noise (having the p-values $\alpha = 0.003$, $\beta < 0.001$ and Shapiro-Wilk = 0.383) and that a_t from Eq.(12) does not contain significant heteroscedasticity except that ϵ_t is an i.i.d. white noise with a heavy-tailed distribution (having the p-values $\alpha = 0.136$, $\beta = 0.454$ and Shapiro-Wilk = 0.018). We obtain GARCH(1,1)-filtered residuals and the test statistic in Eq.(8) is re-applied to three sets of residuals: OLS residuals from Eq.(11) and Eq.(12); GARCH(1,1)-filtered residuals of stock returns and OLS residuals from Eq.(12); and GARCH(1,1)-filtered residuals from both stock returns and Anxiety Index. The results we present in rows 2 and 3 of Table 3 show p-values > 0.05 and thus confirm that our earlier result presented in row 1 of Table 3 is biased by the presence of heteroscedasticity in the residuals. We are thus able to show that the Anxiety Index does not possess any significant predictive information on the stock market.

In view of our results above, we therefore claim that the conclusion from Gilbert and Karahalios [8] according to which the Anxiety Index has predictive information on the stock market is not valid, which is supported also by the fact that the statistical conditions to validate their results are not met.

4 Conclusion

This paper proposes a new approach to statistically demonstrating the predictive information relationship direction between stock prices and emotions expressed

Table 3. Non-linear Granger non-causality test

$AI => SP$		$SP => AI$	
Lx=Ly=1	p	**Lx=Ly=1**	p
Before filtering	0.017	Before filtering	0.182
$GARCH(1,1)_{SP}$	0.349	$GARCH(1,1)_{SP}$	0.922
$GARCH(1,1)_{SP,AI}$	0.718	$GARCH(1,1)_{SP,AI}$	0.685

online. In particular it proves that the Anxiety Index introduced by Gilbert and Karahalios [8] does not possess predictive information with respect to S&P 500. Our work does so by addressing the statistical limitations present in, and by extending the approach of [8].

The main drawback of the approach in [8] to proving the existence of the predictive information of the Anxiety Index with respect to the stock market was that this approach used a Granger causality analysis based on producing and assessing predictive linear models, which were actually not valid from a statistical point of view. In particular these models suffered of major shortcomings as for instance residuals were non-normally distributed, and they presented a heterogeneity of the variance. In an attempt to partially correct the above shortcomings, the Monte Carlo simulation performed by assuming a Gaussian kernel based density for the Anxiety Index, was also biased as the empirical distribution of the employed F statistic significantly deviated from the expected F-distribution [8].

We note that Monte Carlo simulations using the Gaussian kernel density approach have their own bandwidth selection problem, which may bias the simulations - see Zambom and Dias [17]. We therefore re-designed the Monte Carlo simulation presented in [8] by using bootstrap samples of the Anxiety Index first, and the inverse transform sampling based on the continuous version of the empirical distribution function corresponding to the original Anxiety Index sample. The results showed no improvement. This re-confirms the non-linear nature in the relationship between the stock market and emotion, and the erratic volatility in the variables. Linear models appear to be too 'basic' to capture these complexities.

We have therefore extended the approach of [8] by proposing a more capable framework based on the non-linear models introduced in [6]. Our first result, based on a p-value of 0.017 obtained in the non-linear Granger non-causality test, capturing the predictive information of the Anxiety Index with respect to S&P 500, is biased by the presence of heteroscedasticity. We filtered out the heteroscedasticity in the residuals using Eq. (13) and our GARCH(1,1)-filtered residuals were used with the test statistic in Eq. (7). Our results, based on p-values > 0.05, express the true non-causality relationship of Anxiety Index with respect to S&P 500.

Although our work has established that the Anxiety Index does not have predictive information with respect to the stock market, by proposing a new

approach which is statistically sound and more conclusive, there are still some concerns on how the Anxiety Index was built, based on incomplete data, non-specific LiveJournal posts, corpus challenges, non-representative data sample, among others. Further refining the process of defining the Anxiety Index by addressing the above mentioned concerns, may help to fine-tune our empirical results and provide us with a more reliable predictive model.

References

[1] Anderson T. G.: Return volatility and trading volume: an information flow interpretation of stochastic volatility. *Journal of Finance,* **51**, 169–204, (1996).

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

[3] Balog K., Gilad M., Maarten de R.: 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).

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

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

[6] 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).

[7] *http://www1.fee.uva.nl/cendef/whoiswho/showHP/default.asp?selected=40&pid=6*

[8] 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).

[9] 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).

[10] Jahidul A., Mohammad A.H., Rajib H.: Analyzing public emotion and predicting stock market using social media. *American Journal of Engineering Research,* **2(9)**, 265–275, (2013).

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

[12] Mao H., Counts A., Bollen J.: Predicting financial markets: comparing survey, news, twitter and search engine data. *arXiv preprint,* **arXiv**:1112.1051 (2011).

[13] Mittal A., Goel A.: Stock prediction using twitter sentiment analysis, *Project report, Stanford,* (2012).

[14] Schumaker R.P., Chen H.: 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).

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

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

[17] Zambom A.Z., Dias R.: A review of kernel density estimation with application to Econometrics. *arXiv:1212.2812v1*, (2012)