Twitter mood predicts the stock market

Johan Bollen [a],[*],[1], Huina Mao [a],[1], Xiaojun Zeng [b]

[a] *School of Informatics and Computing, Indiana University, 919 E. 10th Street, Bloomington, IN 47408, United States*
[b] *School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom*

ABSTRACT

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error (MAPE) by more than 6%.

1. Introduction

Stock market prediction has attracted much attention from academia as well as business. But can the stock market really be predicted? Early research on stock market prediction [14,13,19] was based on random walk theory and the Efficient Market Hypothesis (EMH) [12]. According to the EMH stock market prices are largely driven by *new* information, i.e. news, rather than present and past prices. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted with more than 50% accuracy [43].

A growing body of research has however critically examined EMH [31], in particular from the perspective of the Socionomic Theory of Finance (STF) [42,41], behavioral economics [47] and behavioral finance [36]. Numerous studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted [4,24,16,43] thereby calling into question EMH's basic assumptions. Some recent research also suggests that news may be unpredictable but that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc.) to predict changes in various economic and commercial indicators. This may conceivably also be the case for the stock market. For example, Gruhl et al. [18] showed how online chat activity predicts book sales. Mishne and Rijke [34] used assessments of blog sentiment to predict movie sales. Liu et al. [30] predicted the future product sales using a Probabilistic Latent Semantic Analysis (PLSA) model to extract indicators of sentiment from blogs. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [6]. Schumaker and Chen [46] investigated the relations between breaking financial news and stock price changes. Most recently, Asur and Huberman [1] provided a demonstration of how public sentiment related to movies, as expressed on Twitter, can actually predict box office receipts.

Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play an significant role in human decision-making [9,7,23]. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [36]. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news.

However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for

* Corresponding author. Tel.: +1 812 856 1833.
  E-mail addresses: jbollen@indiana.edu (J. Bollen), huinmao@indiana.edu
(H. Mao), x.zeng@manchester.ac.uk (X. Zeng).
  [1] Authors made equal contributions.

practical stock market prediction. Large surveys of public mood over representative samples of the population are generally expensive and time-consuming to conduct, cf. Gallup's opinion polls and various consumers and well-being indices. Some have therefore proposed indirect assessment of public mood or sentiment from the results of soccer games [10] and from weather conditions [20]. The accuracy of these methods is however limited by the low degree to which the chosen indicators are expected to be correlated with public mood.

Over the past 5 years significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content [17,34,30,8] and in particular large-scale Twitter feeds [39]. Although each so-called *tweet*, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to Twitter at any given time may provide an accurate representation of public mood and sentiment. This has led to the development of real-time sentiment-tracking indicators such as [8] and "Pulse of Nation".[2] In fact, recent work has started to explore the extraction of such indicators of public mood state from online sources and relating them to economic indicators [17,50].

In this paper we investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can indeed be used to predict the stock market. We use two tools to measure variations in the public mood from tweets submitted to the Twitter service from February 28, 2008 to December 19, 2008. The first tool, OpinionFinder, analyses the text content of tweets submitted on a given day to provide a positive vs. negative daily time series of public mood. The second tool, GPOMS, similarly analyses the text content of tweets to generate a six-dimensional daily time series of public mood to provide a more detailed view of changes in public along a variety of different mood dimensions. The resulting public mood time series are correlated to the Dow Jones Industrial Average (DJIA) to assess their ability to predict changes in the DJIA over time. Our results indicate that the prediction accuracy of standard stock market prediction models is significantly improved when certain mood dimensions are included, but not others. In particular variations along the public mood dimensions of Calm and Happiness as measured by GPOMS seem to have a predictive effect, but not general happiness as measured by the OpinionFinder tool.

2. Results

2.1. Data and methods overview

We obtained a collection of public tweets that was recorded from February 28 to December 19th, 2008 (9,853,498 tweets posted by approximately 2.7 M users). For each tweet these records provide a tweet identifier, the date–time of the submission (GMT+0), its submission type, and the text content of the Tweet which is by design limited to 140 characters. After removal of stop-words and punctuation, we group all tweets that were submitted on the same date. We only take into account tweets that contain explicit statements of their author's mood states, i.e. those that match the expressions "i feel","i am feeling","i'm feeling","i dont feel", "I'm", "Im", "I am", and "makes me". In order to avoid spam messages and other information-oriented tweets, we also filter out tweets that match the regular expressions "http:" or "www."

As shown in Fig. 1 we then proceed in three phases. In the first phase, we subject the collections of daily tweets to 2 mood assessment tools: (1) OpinionFinder which measures positive vs. negative

mood from text content, and (2) GPOMS which measures 6 different mood dimensions from text content. This results in a total of 7 public mood time series, one generated by OpinionFinder and six generated by GPOMS, each representing a potentially different aspects of the public's mood on a given day. In addition, we extract a time series of daily DJIA closing-values from Yahoo! Finance. In the second phase, we investigate the hypothesis that public mood as measured by GPOMS and OpinionFinder is predictive of future DJIA values. We use a Granger causality analysis in which we correlate DJIA values to GPOMS and OF values of the past *n* days. In the third phase, we deploy a Self-Organizing Fuzzy Neural Network model to test the hypothesis that the prediction accuracy of DJIA prediction models can be improved by including measurements of public mood. We are not interested in proposing an optimal DJIA prediction model, but to assess the effects of including public mood information on the accuracy of a "baseline" prediction model.

2.2. Generating public mood time series: OpinionFinder and GPOMS

OpinionFinder (OF)[3] is a publicly available software package for sentiment analysis that can be applied to determine sentence-level subjectivity [48], i.e. to identify the emotional polarity (positive or negative) of sentences. It has been successfully used to analyze the emotional content of large collections of tweets [38] using the OF lexicon to determine the ratio of positive versus negative tweets on a given day. The resulting time series were shown to correlate with the Consumer Confidence Index from Gallup[4] and the Reuters/University of Michigan Surveys of Consumers[5] over a given period of time. We adopt OF's subjective lexicon that has been established upon previous work [45,44,48]. We select positive and negative words that are marked as either "weak" and "strong" from the OF sentiment lexicon resulting in a list of 2718 positive and 4912 negative words. For each tweet we determine whether it contains any number of negative and positive terms from the OF lexicon. For each occurrence we increase the score of either negative or positive tweets by 1 and calculate the ratio of positive vs. negative messages for the tweets posted on the same day *t*.

Like many sentiment analysis tools OF adheres to a unidimensional model of mood, making binary distinctions between positive and negative sentiment [40]. This may however ignore the rich, multi-dimensional structure of human mood. To capture additional dimensions of public mood we created a second mood analysis tools, labeled GPOMS, that can measure human mood states in terms of 6 different mood dimensions, namely *Calm*, *Alert*, *Sure*, *Vital*, *Kind* and *Happy*. GPOMS' mood dimensions and lexicon are derived from an existing and well-vetted psychometric instrument, namely the Profile of Mood States (POMS-bi) [37,33]. To make it applicable to Twitter mood analysis we expanded the original 72 terms of the POMS questionnaire to a lexicon of 964 associated terms by analyzing word co-occurrences in a collection of 2.5 billion 4- and 5-grams[6] computed by Google in 2006 from approximately 1 trillion word tokens observed in publicly accessible Webpages [3,2]. The enlarged lexicon of 964 terms thus allows GPOMS to capture a much wider variety of naturally occurring mood terms in Tweets and map them to their respective POMS mood dimensions. We match the terms used in each tweet against this lexicon. Each tweet term that matches an *n*-gram term is mapped back to its
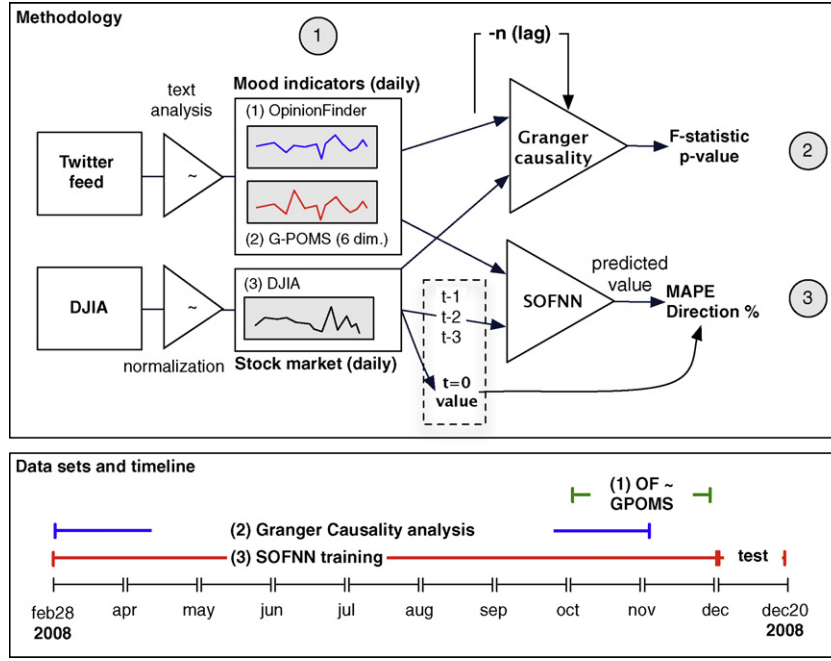
---

[2] See http://www.ccs.neu.edu/home/amislove/twittermood/.

[3] See http://www.cs.pitt.edu/mpqa/opinionfinderrelease/.
[4] See http://www.gallup.com/poll/122840/Gallup-Daily-Economic-Indexes.aspx.
[5] See http://www.sca.isr.umich.edu/.
[6] n-grams are frequently occurring sequences of terms in text of length *n*, for example "we are the robots" could be a frequent 4-gram.

Fig. 1. Diagram outlining 3 phases of methodology and corresponding data sets: (1) creation and validation of OpinionFinder and GPOMS public mood time series from October 2008 to December 2008 (Presidential Election and Thanksgiving), (2) use of Granger causality analysis to determine correlation between DJIA, OpinionFinder and GPOMS public mood from August 2008 to December 2008, and (3) training of a Self-Organizing Fuzzy Neural Network to predict DJIA values on the basis of various combinations of past DJIA values and OF and GPOMS public mood data from March 2008 to December 2008.

original POMS terms (in accordance with its co-occurence weight) and via the POMS scoring table to its respective POMS dimension. The score of each POMS mood dimension is thus determined as the weighted sum of the co-occurence weights of each tweet term that matched the GPOMS lexicon. Data sets and methods are available on our project web site.[7]

To enable the comparison of OF and GPOMS time series we normalize them to z-scores on the basis of a local mean and standard deviation within a sliding window of k days before and after the particular date. For example, the z-score of time series X_t, denoted $\mathbb{Z}_{X_t}$, is defined as:

$$\mathbb{Z}_{X_t} = \frac{X_t - \bar{x}(X_{t\pm k})}{\sigma(X_{t\pm k})} \tag{1}$$

where $\bar{x}(X_{t\pm k})$ and $\sigma(D_{t\pm k})$ represent the mean and standard deviation of the time series within the period $[t-k, t+k]$. This normalization causes all time series to fluctuate around a zero mean and be expressed on a scale of 1 standard deviation.

The mentioned z-score normalization is intended to provide a common scale for comparisons of the OF and GPOMS time series. However, to avoid so-called "in-sample" bias, we do *not* apply z-score normalization to the mood and DJIA time series that are used to test the prediction accuracy of our Self-Organizing Fuzzy Neural Network in Section 2.5. This analysis and our prediction results rest on the raw values for both time series and the DJIA.

2.3. Cross-validating OF and GPOMS time series against large socio-cultural events

We first validate the ability of OF and GPOMS to capture various aspects of public mood. To do so we apply them to tweets posted in a 2-month period from October 5, 2008 to December 5, 2008. This period was chosen specifically because it includes several socio-cultural events that may have had a unique, significant and complex

effect on public mood namely the U.S presidential election (November 4, 2008) and Thanksgiving (November 27, 2008). The OF and GPOMS measurements can therefore be cross-validated against the expected emotional responses to these events. The resulting mood time series are shown in Fig. 2 and are expressed in z-scores as given by in Eq. (1).
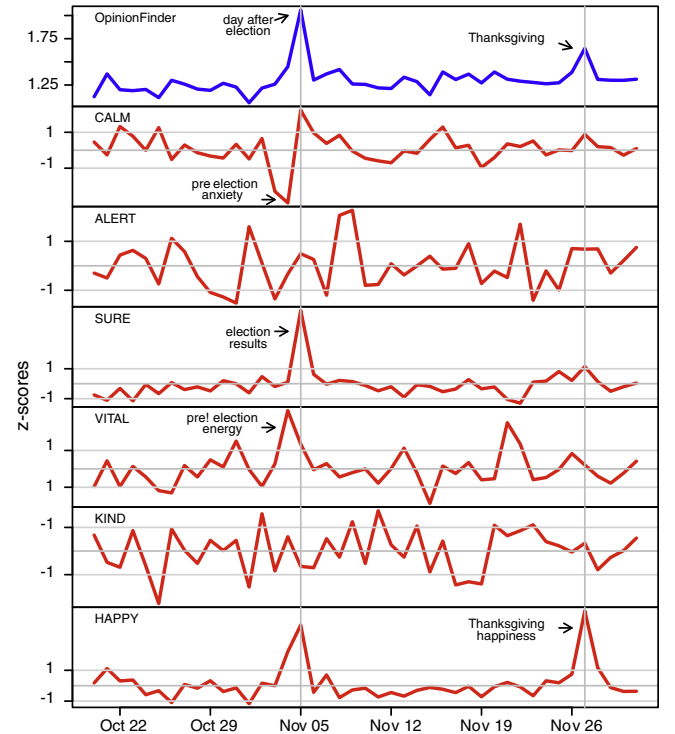
Fig. 2. Tracking public mood states from tweets posted between October 2008 to December 2008 shows public responses to presidential election and thanksgiving.

Table 1
Multiple regression results for OpinionFinder vs. 6 GPOMS mood dimensions.

Parameters	Coeff.	Std. Err.	t	P
Calm (X_1)	1.731	1.348	1.284	0.205
Alert (X_2)	0.199	2.319	0.086	0.932
Sure (X_3)	3.897	0.613	6.356	**4.25e−08**[***]
Vital (X_4)	1.763	0.595	2.965	**0.004**[*]
Kind (X_5)	1.687	1.377	1.226	0.226
Happy (X_6)	2.770	0.578	4.790	**1.30e−05**[**]
Summary	Residual Std. Err.	Adj. R^2	$F_{6,55}$	p
	0.078	0.683	22.93	2.382e−13

[*] $p < 0.1$.
[**] $p < 0.05$.
[***] $p < 0.001$.

Table 2
Statistical significance (p-values) of bivariate Granger-causality correlation between moods and DJIA in period February 28, 2008 to November 3, 2008.

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	**0.085**[*]	0.272	0.952	0.648	0.120	0.848	0.388
2 Days	0.268	**0.013**[**]	0.973	0.811	0.369	0.991	0.7061
3 Days	0.436	**0.022**[**]	0.981	0.349	0.418	0.991	0.723
4 Days	0.218	**0.030**[**]	0.998	0.415	0.475	0.989	0.750
5 Days	0.300	**0.036**[**]	0.989	0.544	0.553	0.996	0.173
6 Days	0.446	**0.065**[*]	0.996	0.691	0.682	0.994	**0.081**[*]
7 Days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

[*] $p < 0.1$.
[**] $p < 0.05$.

Fig. 2 shows that the OF successfully identifies the public's emotional response to the Presidential election on November 4th and Thanksgiving on November 27th. In both cases OF marks a significant, but short-lived uptick in positive sentiment specific to those days.

The GPOMS results reveal a more differentiated public mood response to the events in the 3-day period surrounding the election day (November 4, 2008). November 3, 2008 is characterized by a significant drop in Calm indicating highly elevated levels of public anxiety. Election Day itself is characterized by a reversal of Calm scores indicating a significant reduction in public anxiety, in conjunction with a significant increases of Vital, Happy as well as Kind scores. The latter indicates a public that is energized, happy and friendly on election day. On November 5, these GPOMS dimensions continue to indicate positive mood levels, in particular high levels of Calm, Sure, Vital and Happy. After November 5, all mood dimensions gradually return to the baseline. The public mood response to Thanksgiving on November 27, 2008 provides a counterpart to the differentiated response to the Presidential Election. On Thanksgiving day we find a spike in Happy values, indicating high levels of public happiness. However, no other mood dimensions are elevated on November 27. Furthermore, the spike in Happy values is limited to the 1 day, i.e. we find no significant mood response the day before or after Thanksgiving.

A visual comparison of Fig. 2 suggests that GPOMS' Happy dimension best approximates the mood trend provided by OpinionFinder. To quantitatively determine the relations between GPOMS's mood dimensions and the OF mood trends, we test the correlation between the trend obtained from OF lexicon and the six dimensions of GPOMS using multiple regression. The regression model is shown in Eq. (2).

$$Y_{OF} = \alpha + \sum_{i=1}^{N} \beta_i X_i + \varepsilon_t \tag{2}$$

where $N = 6$, X_1, X_2, X_3, X_4, X_5, and X_6 represent the mood time series obtained from the 6 GPOMS dimensions, respectively Calm, Alert, Sure, Vital, Kind and Happy.

The multiple linear regression results are provided in Table 1 (coefficient and p-values), and indicate that Y_{OF} is significantly correlated with X_3 (Sure), X_4 (Vital) and X_6 (Happy), but not with X_1 (Calm), X_2 (Alert) and X_5 (Kind). We therefore conclude that certain GPOMS mood dimension partially overlap with the mood values provided by OpinionFinder, but not necessarily all mood dimensions that may be important in describing the various components of public mood e.g. the varied mood response to the Presidential Election. The GPOMS thus provides a unique perspective on public mood states not captured by uni-dimensional tools such as OpinionFinder.

2.4. Bivariate Granger causality analysis of mood vs. DJIA prices

After establishing that our mood time series responds to significant socio-cultural events such as the Presidential Election and Thanksgiving, we are concerned with the question whether other variations of the public's mood state correlate with changes in the stock market, in particular DJIA closing values. To answer this question, we apply the econometric technique of Granger causality analysis to the daily time series produced by GPOMS and OpinionFinder vs. the DJIA. Granger causality analysis rests on the assumption that if a variable X causes Y then changes in X will systematically occur before changes in Y. We will thus find that the lagged values of X will exhibit a statistically significant correlation with Y. Correlation however does not prove causation. We therefore use Granger causality analysis in a similar fashion to [17]; we are not testing actual causation but whether one time series has predictive information about the other or not.[8]

Our DJIA time series, denoted D_t, is defined to reflect daily changes in stock market value, i.e. its values are the delta between day t and day $t-1$: $D_t = DJIA_t - DJIA_{t-1}$. To test whether our mood time series predicts changes in stock market values we compare the variance explained by two linear models as shown in Eqs. (3) and (4). The first model (L_1) uses only n lagged values of D_t, i.e. ($D_{t-1}, \cdots, D_{t-n}$) for prediction, while the second model L_2 uses the n lagged values of both D_1 and the GPOMS plus the OpinionFinder mood time series denoted $X_{t-1}, \cdots, X_{t-n}$.

We perform the Granger causality analysis according to model L_1 and L_2 shown in Eqs. (3) and (4) for the period of time between February 28 to November 3, 2008 to exclude the exceptional public mood response to the Presidential Election and Thanksgiving from the comparison. GPOMS and OpinionFinder time series were produced for 342,255 tweets in that period, and the daily Dow Jones Industrial Average (DJIA) was retrieved from Yahoo! Finance for each day.[9]

$$L_1 : \quad D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-i} + \varepsilon_t \tag{3}$$

$$L_2 : \quad D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \varepsilon_t \tag{4}$$

Based on the results of our Granger causality (shown in Table 2), we can reject the null hypothesis that the mood time series do not predict DJIA values, i.e. $\beta_{\{1,2,\cdots,n\}} \neq 0$ with a high level of confidence. However, this result only applies to 1 GPOMS mood

[8] Gilbert and Karahalios [17] uses only one mood index, namely Anxiety, but we investigate the relation between DJIA values and all Twitter mood dimensions measured by GPOMS and OpinionFinder.
[9] Our DJIA time series has no values for weekends and holidays because trading is suspended during those days. We do not linearly extropolate to fill the gaps. This results in a time series of 64 days.
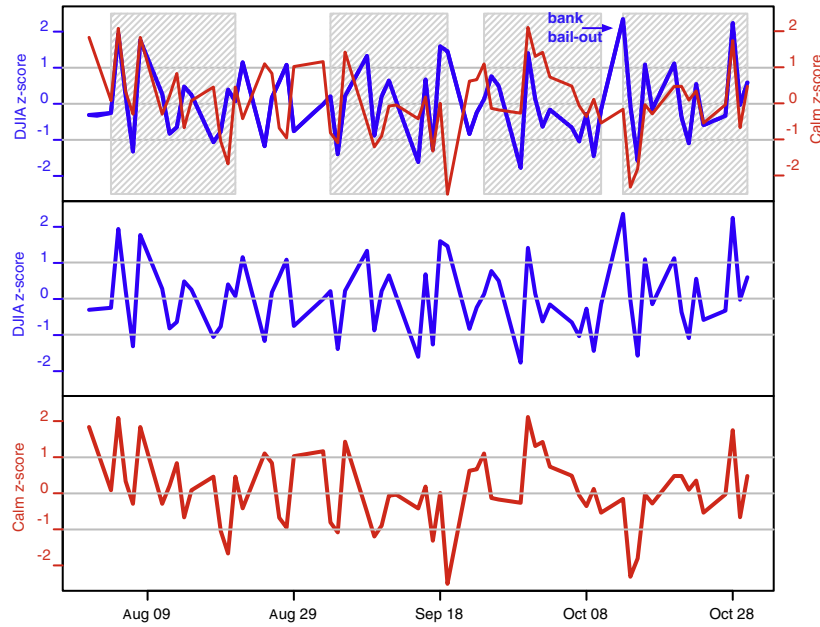
Fig. 3. A panel of three graphs. The top graph shows the overlap of the day-to-day difference of DJIA values (blue : $\mathbb{Z}_{D_t}$) with the GPOMS' Calm time series (red : $\mathbb{Z}_{X_t}$) that has been lagged by 3 days. Where the two graphs overlap the Calm time series predict changes in the DJIA closing values that occur 3 days later. Areas of significant congruence are marked by gray areas. The middle and bottom graphs show the separate DJIA and GPOMS' Calm time series. (For interpretation of the references to color in text, the reader is referred to the web version of the article.)

dimension. We observe that X_1 (i.e. Calm) has the highest Granger causality relation with DJIA for lags ranging from 2 to 6 days (p-values < 0.05). The other four mood dimensions of GPOMS do not have significant causal relations with changes in the stock market, and neither does the OpinionFinder time series.

To visualize the correlation between X_1 and the DJIA in more detail, we plot both time series in Fig. 3. To maintain the same scale, we convert the DJIA delta values D_1 and mood index value X_1 to z-scores as shown in Eq. (1).

As can be seen in Fig. 3 both time series frequently overlap or point in the same direction. Changes in past values of Calm ($t-3$) predicts a similar rise or fall in DJIA values ($t-0$). The Calm mood dimension thus has predictive value with regards to the DJIA. In fact the p-value for this shorter period, i.e. August 1, 2008 to October 30, 2008, is significantly lower (lag $n-3$, $p = 0.009$) than that listed in Table 2 for the period February 28, 2008 to November 3, 2008.

The cases in which the $t-3$ mood time series fails to track changes in the DJIA are nearly equally informative as where it does not. In particular we point to a significant deviation between the two graphs on October 13th where the DJIA surges by more than 3 standard deviations trough-to-peak. The Calm curve however remains relatively flat at that time after which it starts to again track changes in the DJIA again. This discrepancy may be the result of the the Federal Reserve's announcement on October 13th of a major bank bailout initiative which unexpectedly increase DJIA values that day. The deviation between Calm values and the DJIA on that day illustrates that unexpected news is not anticipated by the public mood yet remains a significant factor in modeling the stock market.

2.5. Non-linear models for emotion-based stock prediction

Our Granger causality analysis suggests a predictive relation between certain mood dimensions and DJIA. However, Granger causality analysis is based on linear regression whereas the relation between public mood and stock market values is almost certainly non-linear. To better address these non-linear effects and assess the contribution that public mood assessments can make in pre-

dictive models of DJIA values, we compare the performance of a Self-organizing Fuzzy Neural Network (SOFNN) model [28] that predicts DJIA values on the basis of two sets of inputs: (1) the past 3 days of DJIA values, and (2) the same combined with various permutations of our mood time series (explained below). Statistically significant performance differences will allow us to either confirm or reject the null hypothesis that public mood measurement do not improve predictive models of DJIA values.

Neural networks have previously been used to decode non-linear time series data which describe the characteristics of the stock market [26] and predict stock market values [51,25]. SOFNN combines the learning ability of neural networks with the easy interpretability of fuzzy systems. Whereas popular self-organizing neural networks such as Grossberg's ART [5], Nigrin's SONNET [35] and Hopfield network [21] were originally developed for pattern classification problems, SOFNN has been developed specifically for regressions, function approximation and time series analysis problems. Compared with some notable fuzzy nerural network models, such as the adaptive-network-based fuzzy inference systems (ANFIS) [22], self-organizing dynamic fuzzy neural network (DFNN) [11] and GDFNN [49], SOFNN provides a more efficient algorithm for online learning due to its simple and effective parameter and structure learning algorithm [28]. In our previous work, SOFNN has proven its value in electrical load forecasting [32], exchange rate forecasting [28] and other applications [29].

To predict the DJIA value on day t, the input attributes of our SOFNN include combinations of DJIA values and raw mood values of the past n days (not normalized to z-scores). We choose $n-3$ since the results shown in Table 2 indicate that past $n-3$ the Granger causal relation between Calm and DJIA decreases significantly. All historical load values are linearly scaled to [0,1]. This procedure causes every input variable be treated with similar importance since they are processed within a uniform range.

SOFNN models require the tuning of a number of parameters that can influence the performance of the model. We maintain the same parameter values across our various input combinations to allow an unbiased comparison of model performance, namely $\delta = 0.04$, $\sigma = 0.01$, $k_{rmse} = 0.05$, $k_d(i)$, ($i = 1, \ldots, r$) = 0.1 where r is the
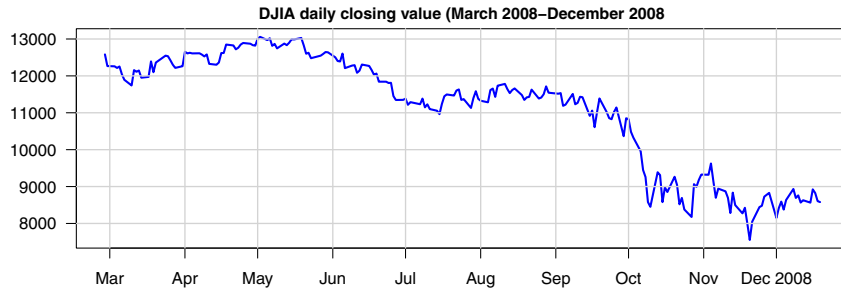
Fig. 4. Daily Dow Jones Industrial Average values between February 28, 2008 and December 19, 2008.

dimension of input variables and k_{rmse} is the expected training root mean squared error which is a predefined value.

To properly evaluate the SOFNN model's ability to predict daily DJIA prices, we extend the period under consideration to February 28, 2008 to December 19, 2008 for training and testing. February 28, 2008 to November 28, 2008 is chosen as the longest possible training period while December 1 to December 19, 2008 was chosen as the test period because it was characterized by stabilization of DJIA values after considerable volatility in previous months and the absence of any unusual or significant socio-cultural events. Fig. 4 shows that the Fall of 2008 is an unusual period for the DJIA due to a sudden dramatic decline of stock prices. This variability may in fact render stock market prediction more difficult than in other periods.

The Granger causality analysis indicates that only Calm (and to some degree Happy) is Granger-causative of DJIA values. However, the other mood dimensions could still contain predictive information of DJIA values when combined with other mood dimensions. For example, Happy may not be independently linearly related with DJIA, but it may nevertheless improve the SOFNN prediction accuracy when combined with Calm. To clarify these questions, we investigate seven permutations of input variables to the SOFNN model, the first of which, denoted I_0, represents a naive, baseline model that has been trained to predict DJIA values at time t from the historical values at time $\{t-1,t-2,t-3\}$:

$$I_0 = \{DJIA_{t-3,2,1}\}$$
$$I_1 = \{DJIA_{t-3,2,1}, X_{1,t-3,2,1}\}$$
$$I_{1,2} = \{DJIA_{t-3,2,1}, X_{1,t-3,2,1}, X_{2,t-3,2,1}\}$$
$$I_{1,3} = \{DJIA_{t-3,2,1}, X_{1,t-3,2,1}, X_{3,t-3,2,1}\}$$
$$\cdots$$

$DJIA_{t-3,2,1}$ represents the DJIA values and $X_{1,t-3,2,1}$ represents the values of the GPOMS mood dimension 1, at time $t-3$, $t-2$, and $t-1$. According to the same notation $I_{1,3}$, $I_{1,4}$, $I_{1,5}$, $I_{1,6}$ represent a combination of historical DJIA with mood dimensions 3, 4, 5 and 6 at time $t-3$, $t-2$, and $t-1$. For example, $I_{1,6}$ represents a set of inputs that includes the DJIA values $t-3$, $t-2$, and $t-1$, and mood dimensions 1 and 6 at the same times.

In order to compare the performance of the GPOMS mood data with the positive/negative sentiment values given by Opinion-Finder, we additionally define the input combination:

$$I_{OF} = \{DJIA_{t-3,2,1}, X_{OF,t-3,2,1}\}$$

Forecasting accuracy is measured in terms of the average Mean Absolute Percentage Error (MAPE) and the direction accuracy (up or down) during the test period (December 1 to December 19, 2008). The prediction results are shown in Table 3.

We can draw several conclusions from these results. First, adding positive/negative sentiment obtained from OF (I_{OF}) has no effect on prediction accuracy compared to using only historical DJIA values (I_0). This confirms the results of our Granger causuality analysis.

Second, adding Calm, i.e. input I_1, we find the highest prediction accuracy. Compared to I_0 and all other input combinations, adding input I_1 leads to significant improvements in MAPE values (1.83% vs. the maximum of 2.13% and 1.95% for I_{OF}) and direction accuracy (86.7% compared to 73.3% for I_{OF} and 46.7% for $I_{1,3}$). Thirdly, $I_{1,3}$ and $I_{1,4}$ actually reduce prediction accuracy significantly in terms of both MAPE and direction %, most likely because X_3 (Sure) and X_4 (Vital) do not contain information that is useful for prediction of DJIA values causing prediction accuracy to regress to chance levels. It is notable that $I_{1,6}$, i.e. a combination of X_6 and X_1 does significantly reduce average MAPE, and provides good direction accuracy (80%). This is surprising since X_6 (i.e. Happy) does not have a good Granger causality relation with DJIA at a lag of $n = 3$ days (see Table 2, $p = 0.723$). However in combination with Calm, it produces a more accurate SOFNN prediction (MAPE = 1.79%) and direction accuracy (80%).

To assess the statistical significance of the SOFNN achieving the above mentioned accuracy of 87.6% in predicting the up and down movement of the DJIA we calculate the odds of this result occurring by chance. The binomial distribution indicates that the probability of achieving exactly 87.6% correct guesses over 15 trials (20 days minus weekends) with a 50% chance of success on each single trial equals 0.32%. Taken over the entire length of our data set (February 28 to December 20, excluding weekends) we find approximately 10.9 of such 20-day periods. The odds that the mentioned probability would hold by chance for a random period of 20 days within that period is then estimated to be $1 - (1 - 0.0032)^{10.9} = 0.0343$ or 3.4%. The SOFNN direction accuracy is thus most likely not the result of chance nor our selecting a specifically favorable test period.

In addition, we test the linear effect of both Calm (X_1) and Happy (X_6) on DJIA, with a nested F-test between the full model F and reduced model R shown as follows:

$$\mathbf{F}: \quad D_t = \alpha + \sum_{i=1}^{n}\beta_i D_{t-i} + \sum_{i=1}^{3}\gamma_i X_{1,t-i} + \sum_{i=1}^{3}\gamma_i X_{6,t-i} + \varepsilon_t$$

$$\mathbf{R}: \quad D_t = \alpha + \sum_{i=1}^{n}\beta_i D_{t-i} + \sum_{i=1}^{3}\gamma_i X_{1,t-i} + \varepsilon_t$$

Table 3
DJIA daily prediction using SOFNN.

Evaluation	I_{OF}	I_0	I_1	$I_{1,2}$	$I_{1,3}$	$I_{1,4}$	$I_{1,5}$	$I_{1,6}$
MAPE (%)	1.95	1.94	1.83	2.03	2.13	2.05	1.85	**1.79***
Direction (%)	73.3	73.3	**86.7**	60.0	46.7	60.0	73.3	80.0

We find a p-value of 0.66 and an F-statistic of 0.53 indicating that a linear combination of X_1 and X_6 produces worse results than X_1 alone. Since the SOFNN prediction is more accurate when using a combination of X_1 and X_6, we conclude that this confirms a non-linear relation among the different dimensions of moods.

3. Discussion

In this paper, we investigate whether public mood as measured from large-scale collection of tweets posted on twitter.com is correlated or even predictive of DJIA values. Our results show that changes in the public mood state can indeed be tracked from the content of large-scale Twitter feeds by means of rather simple text processing techniques and that such changes respond to a variety of socio-cultural drivers in a highly differentiated manner. Among the 7 observed mood dimensions only some are Granger causative of the DJIA; changes of the public mood along these mood dimensions match shifts in the DJIA values that occur 3–4 days later. Surprisingly we do not observe this effect for OpinionFinder's assessment of public mood states in terms of positive vs. negative mood but rather for the GPOMS dimension labeled "Calm". The calmness of the public (measured by GPOMS) is thus predictive of the DJIA rather than general levels of positive sentiment as measured by OpinionFinder. A Self-Organizing Fuzzy Neural Network trained on the basis of past DJIA values and our public mood time series furthermore demonstrated the ability of the latter to significantly improve the accuracy of even the most basic models to predict DJIA closing values. Given the performance increase for a relatively basic model such as the SOFNN we are hopeful to find equal or better improvements for more sophisticated market models that may in fact include other information derived from news sources, and a variety of relevant economic indicators. These results have implications for existing sentiment tracking tools as well as surveys of "self-reported subjective well-being" in which individuals evaluate the extent to which they experience positive and negative affect, happiness, or satisfaction with life [15]. Such surveys are relatively expensive and time-consuming, and may nevertheless not allow the measurement of public mood along mood dimensions that are relevant to assess particular socio-economic indicators. Public mood analysis from Twitter feeds on the other hand offers an automatic, fast, free and large-scale addition to this toolkit that may in addition be optimized to measure a variety of dimensions of the public mood state.

Our analysis does not acknowledge a number of important factors that will however be examined in future research. First, we note that our analysis is not designed to be limited to any particular geographical location nor subset of the world's population. This approach may be appropriate since the US stock markets are affected by individuals worldwide, but for the particular period under observation Twitter.com users were *de facto* predominantly English speaking and located in the US. As Twitter.com's user base becomes increasingly international and the use of smartphones equipped with geo-location increases, future analysis will have to factor in location and language to avoid geographical and cultural sampling errors. Second, although we have cross-validated the results of 2 different tools to assess public mood states, we have no knowledge of the "ground truth" for public mood states nor in fact for the particular subsample of the population represented by the community of Twitter.com users. This problem can only be addressed by research into direct assessments of public mood states vs. those derived from online communities such as Twitter. Third, these results are strongly indicative of a predictive correlation between measurements of the public mood states from Twitter feeds, but offer no information on the causative mechanisms that may connect online public mood states with DJIA values.

The latter remains a crucial area for future research in which the relation between public mood states and information may need to be examined more closely. Information can with great ease and at very short time-scales travel along the ties in an online social networking environment thereby exerting an equalizing effect on investment strategies (and mood) of individual and professional traders alike. On the other hand, rumors and misleading information can spread with equal if not greater efficiency across the same social networking ties as demonstrated by the prevalence of so-called "astroturfing" and "Twitter bombing" campaigns.[10] This puts large groups of individuals that can not rely on the infrastructure of professional traders at a significant disadvantage and may in fact increase market volatility. As result we might see greater overall market stability but one that is punctuated by significant bursts caused by "social contagion" effects. In addition to such contagion effects future research may need to take into account social and cognitive effects in which individual agents are endowed with the ability to learn from past experiences and can adjust their trading behavior accordingly. The investigation of such phenomena in online social networking environments is part of an exciting new research front commonly referred to as "computational social science" [27].

Acknowledgments

References

[1] S. Asur, B.A. Huberman, Predicting the Future with Social Media, 2010, http://www.arxiv.org arXiv:1003.5699v1.
[2] S. Bergsma, L. Dekang, R. Goebel, Web-scale N-gram models for lexical disambiguation, in: Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09), Pasadena, CA, 2009, pp. 1507–1512.
[3] T. Brants, A. Franz, Web 1T 5-gram Version 1, Tech. rep., Linguistic Data Consortium, Philadelphia, 2006.
[4] K.C. Butler, S.J. Malaikah, Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia, Journal of Banking & Finance 16 (1) (1992) 197–210, http://www.econpapers.repec.org/RePEc:eee:jbfina:v:16:y:1992:i:1:p:197-210.
[5] G.A. Carpenter, S. Grossberg, The ART of adaptive pattern recognition by a self-organizing neural network, Computer 21 (March (3)) (1988) 77–88, http://portal.acm.org/citation.cfm?id=47861.47867.
[6] H. Choi, H. Varian, Predicting the Present with Google Trends, Tech. rep., Google, 2009.
[7] A.R. Damasio, Descartes' Error: Emotion Reason, and the Human Brain, Putnam, 1994.
[8] P.S. Dodds, C.M. Danforth, Measuring the happiness of large-scale written expression: songs, blogs, and presidents, Journal of Happiness (July) (2009).
[9] R.J. Dolan, Emotion cognition, and behavior, Science 298 (5596) (2002) 1191–1194, http://www.sciencemag.org/cgi/content/abstract/298/5596/1191.
[10] A. Edmans, D. García, O.Y. Norli, Sports sentiment and stock returns, Journal of Finance 62 (4) (2007) 1967–1998, http://econpapers.repec.org/RePEc:bla:jfinan:v:62:y:2007:i:4:p:1967-1998.
[11] M.J. Er, S. Wu, A fast learning algorithm for parsimonious fuzzy neural systems, Fuzzy Sets and Systems 126 (3) (2002) 337–351, http://www.sciencedirect.com/science/article/B6V05-4550NT7-5/2/86a57269fc988e7347bae0a2af3f696a.
[12] E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34–105, http://dx.doi.org/10.2307/2350752.
[13] E.F. Fama, Efficient capital markets: II, Journal of Finance 46 (5) (1991) 1575–1617, http://econpapers.repec.org/RePEc:bla:jfinan:v:46:y:1991:i:5:p:1575-617.
[14] E.F. Fama, L. Fischer, M.C. Jensen, R. Roll, The adjustment of stock prices to new information, International Economic Review 10 (February (1)) (1969) 1–21, http://ideas.repec.org/a/ier/iecrev/v10y1969i1p1-21.html.
[15] B.S. Frey, Happiness: A Revolution in Economics, MIT Press Books, The MIT Press, June 2008, http://ideas.repec.org/b/mtp/titles/0262062771.html.

---

[10] See http://www.truthy.indiana.edu/.

[16] L.A. Gallagher, M.P. Taylor, Permanent and temporary components of stock prices: evidence from assessing macroeconomic shocks, Southern Economic Journal 69 (2) (2002) 345–362, http://www.jstor.org/stable/1061676.

[17] E. Gilbert, K. Karahalios, Widespread worry and the stock market, in: Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, 2010, pp. 58–65, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1513/1833.

[18] D. Gruhl, R. Guha, R. Kumar, J. Novak, A. Tomkins, The predictive power of online chatter, in: KDD '05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM Press, New York, NY, 2005, pp. 78–87.

[19] H. Cootner P, The Random Character of Stock Market Prices, MIT, 1964.

[20] D. Hirshleifer, T. Shumway, Good day sunshine: stock returns and the weather, Journal of Finance 58 (3) (2003) 1009–1032, http://ideas.repec.org/a/bla/jfinan/v58y2003i3p1009-1032.html.

[21] J.J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, Proceedings of the National Academy of Sciences of the United States of America 81 (May (10)) (1984) 3088–3092, http://www.pnas.org/content/81/10/3088.abstract.

[22] J.S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man and Cybernetics 23 (August (3)) (2002) 665–685, http://dx.doi.org/10.1109/21.256541.

[23] D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk, Econometrica 47 (2) (1979) 263–291.

[24] M. Kavussanos, E. Dockery, A multivariate test for stock market efficiency: the case of ASE, Applied Financial Economics 11 (5) (2001) 573–579, http://econpapers.repec.org/RePEc:taf:apfiec:v:11:y:2001:i:5:p:573-79.

[25] T. Kimoto, K. Asakawa, M. Yoda, M. Takeoka, Stock market prediction system with modular neural networks, in: Proceedings of the International Joint Conference on Neural Networks, IEEE, San Diego, CA, 1990, pp. 1–6.

[26] A. Lapedes, R. Farber, Nonlinear signal processing using neural networks: prediction and system modelling, in: IEEE International Conference on Neural Networks, IEEE, San Diego, CA, 1987.

[27] D. Lazer, D. Brewer, N. Christakis, J. Fowler, G. King, Life in the network: the coming age of computational social science, Science 323 (5915) (2009) 721–723.

[28] G. Leng, G. Prasad, T.M. McGinnity, An on-line algorithm for creating self-organizing fuzzy neural networks, Neural Networks: The Official Journal of the International Neural Network Society 17 (December (10)) (2004) 1477–1493, http://www.ncbi.nlm.nih.gov/pubmed/15541949.

[29] G. Leng, X.-J. Zeng, J.A. Keane, A hybrid learning algorithm with a similarity-based pruning strategy for self-adaptive neuro-fuzzy systems, Applied Soft Computing 9 (September (4)) (2009) 1354–1366, http://portal.acm.org/citation.cfm?id=1595891.1596017.

[30] Y. Liu, X. Huang, A. An, X. Yu, ARSA: a sentiment-aware model for predicting sales performance using blogs, in: SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, 2007, pp. 607–614.

[31] B.G. Malkiel, The efficient market hypothesis and its critics, The Journal of Economic Perspectives 17 (1) (2003), http://dx.doi.org/10.2307/3216840.

[32] H. Mao, X.-J. Zeng, G. Leng, Y. Zhai, J.A. Keane, Short and mid-term load forecasting using a bilevel optimization model, IEEE Transactions On Power Systems 24 (2) (2009) 1080–1090.

[33] D.M. McNair, J.W.P. Heuchert, E. Shilony, Profile of Mood States. Bibliography 1964–2002, Multi-Health Systems, 2003, https://www.mhs.com/ecom/TechBrochures/POMSBibliography.pdf.

[34] G. Mishne, M.D. Rijke, Capturing global mood levels using blog posts, in: N. Nicolov, F. Salvetti, M. Liberman, J.H. Martin (Eds.), AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, The AAAI Press, Menlo Park, CA/Stanford University, CA, August, 2006, pp. 145–152.

[35] A. Nigrin, Neural Networks for Pattern Recognition, MIT Press, Cambridge, MA, 1993.

[36] J.R. Nofsinger, Social mood and financial economics, Journal of Behaviour Finance 6 (3) (2005) 144–160.

[37] J.C. Norcross, E. Guadagnoli, J.O. Prochaska, Factor structure of the profile of mood states (POMS): two partial replications, Journal of Clinical Psychology 40 (5) (2006) 1270–1277.

[38] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May, 2010, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1536.

[39] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, May 2010, pp. 19–21.

[40] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (1–2) (2008) 1–135.

[41] R.R. Prechter, W.D. Parker, The financial/economic dichotomy in social behavioral dynamics: the socionomic perspective, Journal of Behavioral Finance 8 (2) (2007) 84–108, http://dx.doi.org/10.1080/15427560701381028.

[42] R.J. Prechter Robert, The wave principle of human social behavior and the new science of socionomics, New Classics Library, GA, 2002.

[43] Qian, Bo, Rasheed, Khaled, Stock market prediction with multiple classifiers, Applied Intelligence 26 (February (1)) (2007) 25–33, http://dx.doi.org/10.1007/s10489-006-0001-7.

[44] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Morristown, NJ, 2003, pp. 105–112.

[45] E. Riloff, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Association for Computational Linguistics, Morristown, NJ, 2003, pp. 25–32.

[46] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news, ACM Transactions on Information Systems 27 (February (2)) (2009) 1–19, http://portal.acm.org/citation.cfm?doid=1462198.1462204.

[47] V.L. Smith, Constructivist and ecological rationality in economics, American Economic Review 93 (2003) 465–508.

[48] T. Wilson, J. Wiebe, P. Hofmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing—HLT '05 (October), 2005, pp. 347–354, http://portal.acm.org/citation.cfm?doid=1220575.1220619.

[49] S. Wu, M.J. Er, Y. Gao, A fast approach for automatic generation of fuzzy rules by generalized dynamic fuzzy neural networks, IEEE Transactions on Fuzzy Systems 9 (4) (2001) 578–594, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=940970.

[50] X. Zhang, H. Fuehres, P.A. Gloor, Predicting Stock Market Indicators Through Twitter I Hope It is Not as Bad as I Fear, Collaborative Innovation Networks (COINs), Savannah, GA, 2010, http:\\www.elsevier com/locate/procedia.

[51] X. Zhu, H. Wang, L. Xu, H. Li, Predicting stock index increments by neural networks: the role of trading volume under different horizons, Expert Systems with Applications 34 (4) (2008) 3043–3054.
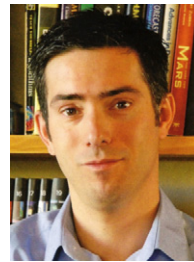
Johan Bollen is associate professor at the Indiana University School of Informatics and Computing. He was formerly a staff scientist at the Los Alamos National Laboratory from 2005-2009, and an Assistant Professor at the Department of Computer Science of Old Dominion University from 2002 to 2005. He obtained his PhD in Experimental Psychology from the University of Brussels in 2001 on the subject of cognitive models of human hypertext navigation. He has taught courses on Data Mining, Information Retrieval and Digital Libraries. His research has been funded by the Andrew W. Mellon Foundation, National Science Foundation, Library of Congress, National Aeronautics and Space Administration and the Los Alamos National Laboratory. His present research interests are sentiment tracking, computational social science, usage data mining, informetrics, and digital libraries. He has extensively published on these subjects as well as matters relating to adaptive information systems architecture. He is presently the Principal Investigator of the Andrew W. Mellon Foundation and NSF funded MESUR project which aims to expand the quantitative tools available for the assessment of scholarly impact.
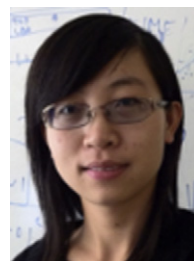
Huina Mao is a Ph.D. student at the Indiana University School of Informatics and Computing since 2008. Her research focus is on electrical load forecasting and web mining. She has published on topics in fuzzy logic, expert systems, and utility load-forecasting. She is presently working on public sentiment tracking from large-scale online, social networking environments with the objective of modeling and predicting socio-economic indicators.

Xiao-Jun Zeng received the B.Sc. degree in Mathematics and the M.Sc. degree in Control Theory and Operation Research from Xiamen University, Xiamen, China, and the Ph.D. degree in Computation from UMIST (currently the University of Manchester), Manchester, U.K., in 1996. He has been with the University of Manchester since 2002 and is currently a Senior Lecturer in the School of Computer Science and the School of Informatics. From 1996 to 2002, he was with Knowledge Support Systems (KSS) Ltd., Manchester, where he was a Scientific Developer, Senior Scientific Developer, and Head of Research, developing intelligent decision support systems which won the European IST Award in 1999 and Microsoft European Retail Application Developer (RAD) Awards in 2001 and 2003. He has published more than 60 journal and conference papers including 12 IEEE Transaction papers. Dr. Zeng is an Associate Editor of the IEEE Trans. on Fuzzy Systems, is a member of the editorial board of the International Journal of Computational Intelligence Research and is an elected member of the EPSRC College.