An entropy based analysis of the relationship between the DOW JONES Index and the TRNA Sentiment series

David E Allen[a], Michael McAleer[b] and Abhay K Singh[c]

[a]Adjunct Professor Centre for Applied Financial Studies, UniSA, and Visiting Professor, School of Mathematics and Statistics, University of Sydney

[b]Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Tinbergen Institute, The Netherlands, Department of Quantitative Economics, Complutense University of Madrid, Spain, and Institute of Economic Research, Kyoto University, Japan

[c]School of Business, Edith Cowan University,Perth, Australia

Abstract

This paper features an analysis of the relationship between the DOW JONES Industrial Average Index (DJIA) and a sentiment news series using daily data obtained from the Thomson Reuters News Analytics (TRNA)[1] provided by SIRCA (The Securities Industry Research Centre of the Asia Pacic). The recent growth in the availability of on-line financial news sources such as internet news and social media sources provides instantaneous access to financial news. Various commercial agencies have started developing their own filtered financial news feeds which are used by investors and traders to support their algorithmic trading strategies. Thomson Reuters News Analytics (TRNA)[2] is one such data set. In this study we use the TRNA data set to construct a series of daily sentiment scores for Dow Jones Industrial Average (DJIA) stock index component companies. We use these daily DJIA market sentiment scores to study the relationship between financial news sentiment scores and the DJIA return series using entropy measures. The entropy and Mutual Information (MI) statistics permit an analysis of the amount of information within the sentiment series and its relationship to the DJIA and an indication of the realtionship changes over time.

Keywords: DJIA, Sentiment, Entropy, TRNA, Information

1. Introduction

The information embodied in news items is one information source that serves to influence investor opinions. The series we use from Thomson Reuters News Analytics (TRNA) could be termed news sentiment and is produced by

Email address: profallen2007@gmail.com (David E Allen[a], Michael McAleer[b] and Abhay K Singh[c])

December 11, 2013

the application of machine learning techniques to news items. These items are calibrated into either positive, negative or neutral values per news item, with implications for the general investor. Investors' investment strategies which influence the market and the evolution of stock prices are potentially influenced by changes in these sentiment stimulated by the continuous flow of news items. Academic researchers and investment practitioners are always looking for new investment tools or factors which may help to predict moves in asset items.

Recently, the role of market news sentiment, in particular machine-driven sentiment signals, and their implication for financial market processes, has been the focus of a great deal of attention. There is a growing body of research that argues that news items from different sources influence investor sentiment, and hence asset prices, asset price volatility and risk (Tetlock, 2007; Telock Saar-Tsechansky, and Macskassy, 2008; Da, Engleberg and Gao, 2011; Odean and Barber, 2008; diBartolomeo and Warrick 2005; Mitra, Mitra and diBartolomeo 2009; Dzielinski, Rieger and Talpsepp 2011). The diversification benefits of the information impounded in news sentiment scores provided by RavenPack has been demonstrated by Cahan, Jussa and Luo (2009) and Hafez and Xie (2012), who examined its benefits in the context of popular asset pricing models.

One important research question is the extent to which the availability of these machine driven series actually contribute to market information and the evolution of security prices. Baker and Wurgler (2006) demonstrated a link between investor sentiment and stock returns. Recent work by Hafez and Xie (2012) examines the effect of investor's sentiment using news based sentiment, generated from the RavenPack Sentiment Index as a proxy for market sentiment in a multi-factor model. They report a strong impact of market sentiment on stock price predictability over 6 and 12 month time horizons.

The issue of the news content of sentiment scores is the central focus of this paper. We address it by analysing the relationship between one commercially available series; the Thomson Reuters News Analytics (TRNA) series and the component stocks of a major index; the DJIA. Given that these large US stocks are likely to be amongst the most heavily traded and analysed securities in the world, the issue of the relevance of these news feeds to their prices and returns series is a central one; with implications for most investments and financial markets.

We take the TRNA news series for the DJIA constituent stocks and aggregate them into a daily time series. This facilitates an analysis of the relationship between the two daily sets of series, TRNA news sentiment on the one hand and DJIA constituent company returns on the other. We analyse the relationship between the two series using entropy based metrics because these are non-parametric and non-linear and should provide a clear indication of the joint information shared by the two sets of series by means of Mutual Information (MI) metrics.

The extent to which these news series have relevant information for security prices and returns is important for both investors and market regulators. If access to these particular information feeds provides a trading advantage, then the market is no longer a level playing field for all investors. Institutions and algorithmic traders with access to these analytics have an advantage. However, this paper does not address the issue of the timing of access to news items, but the more general question of the degree to which these sentiment based series contain 'relevant information'; as revealed by entropy based metrics as applied to the relationship between a daily average TRNA series and daily DJIA returns series.

The paper is organized as follows: Section 1 provides an introduction, Section-2 features an introduction to sentiment analysis and an overview of the TRNA data set and some preliminary statistical analysis. Section-3 discusses entropy metrics and the central research methods used in the empirical exercise undertaken in this paper. The next section-4, discusses the major results and section-5 draws some conclusions.

2. Research methods and data

2.1. News Sentiment

In this paper we examine the sentiment scores provided by TRNA as a single factor using entropy based metrics to evaluate their effect on the stock prices of the DJIA component companies. We use daily DJIA market sentiment scores constructed from high frequency sentiment scores for the various stocks in DJIA. The empirical analysis includes data from the time periods of the Global Financial Crisis and other periods of market turbulence to assess the effect of financial news sentiment on stock prices in both normal and in extreme market conditions. Recently there has been an increase in studies exploring the relationship between stock price movements and news sentiment (Tetlock, 2007, Barber and Odean 2008, Mitra, Mitra and diBartolomeo 2009, Leinweber and Sisk, 2009, Sinha, 2011, Huynh and Smith, 2013).

Given the sheer variety and scale of competing news sources in the electronic media there is scope for the commercial use of sources of pre-processed news. These are available from vendors like TRNA and Ravenpack, who construct sentiment scores to provide direct indicators to traders and other financial practitioners of changes in news sentiment. These sources use text mining tools to electronically analyse available textual news items. The analytics engines of these sources apply pattern recognition and identification methods to analyse, words and their patterns, the novelty and relevance of the news items for a particular industry or sector. The type and characteristics of these news items are converted into quantifiable sentiment scores.

We use sentiment indicators provided by TRNA in our empirical analysis. Thomson Reuters was a pioneer in the implementation of a sophisticated text mining algorithm as an addition to its company and industry specific news database starting from January 2003 which resulted in the present TRNA data set. As per the official TRNA data guide, "Powered by a unique processing system the Thomson Reuters News Analytics system provides real-time numerical insight into the events in the news, in a format that can be directly consumed by algorithmic trading systems". Currently the data set is available for various stocks and commodities until October 2012. The TRNA sentiment scores are produced from text mining news items at a sentence level, which takes into account the context of a particular news item. This kind of news analytics makes the resulting scores more usable as they are mostly relevant to the particular company or sector. Every news item in the TRNA engine is assigned an exact time stamp and a list of companies and topics it mentions. A total of 89 broad fields are reported in the TRNA data set which are broadly divided into following 5 main categories:

1. Relevance: A numerical measure of how relevant the news item is to the asset.
2. Sentiment: A measure of the inherit sentiment of the news item quantifying it as either negative (-1), positive (1) or neutral (0).
3. Novelty: A measure defining how new the news item is; in other words whether it reports a news item that is related to some previous news stories.
4. Volume: Counts of news items related to the particular asset.
5. Headline Classification: Specific analysis of the headline.

Figure-1, shows a snapshot of the headline text as reported in BCAST_REF field of the TRNA database for BHP Billiton during the year 2011. These are not the sentences which are analysed by TRNA to produce sentiment scores but are the headlines for the news item used to generate the TRNA sentiment and other relevant scores. As reported in TRNA, BHP Billiton generated more than 3000 news items in the year 2011. Figure-2 shows the sentiment scores (-1 to +1) for BHP Billiton during the month of January 2011, the red line is the moving average of the scores.

Similar to BHP, there are various news stories reported per day for the various DJIA traded stocks. These news stories result in sentiment scores which are either positive, negative or neutral for that particular stock. Figure-3 gives a snapshot of the sentiment scores for the DJIA's traded stocks during the year 2008. The bar chart of figure-3 shows that the most sentiment scores generated during the year 2008, which was also the period of Global Financial Crisis, were for the Citi Bank group (C.N) , General Motors (GM.N) and J. P Morgan (JPM.N). This is a reflection of the market sentiment during the GFC period, as these financial institutions were among the most effected during the GFC.

Figure-4, shows the number of positive, negative or neutral sentiment scores stacked against each other. Its evident from this figure that the number of negative and neutral sentiment news exceeded the number of positive sentiments for the majority of stocks. Again its in agreement with the context of the GFC period when the DJIA stock market index took a big plunge downwards.

The applications of TRNA news data sets to financial research has recently gained interest. Dzielinski (2012), Groß-Kulßman and Hautsch (2011), Smales (2013), Huynh and Smith (2013), Borovkova and Mahakena (2013). Storkenmaier et al. (2012) and Sinha (2011), have shown the usefulness of the TRNA dataset in stock markets and in commodity markets for both high frequency and multi-day frequency. In this study we utilize the TRNA data set to analyse

Figure 1: TRNA-Snapshot of News Headlines Generated for BHP Billiton in Year 2011

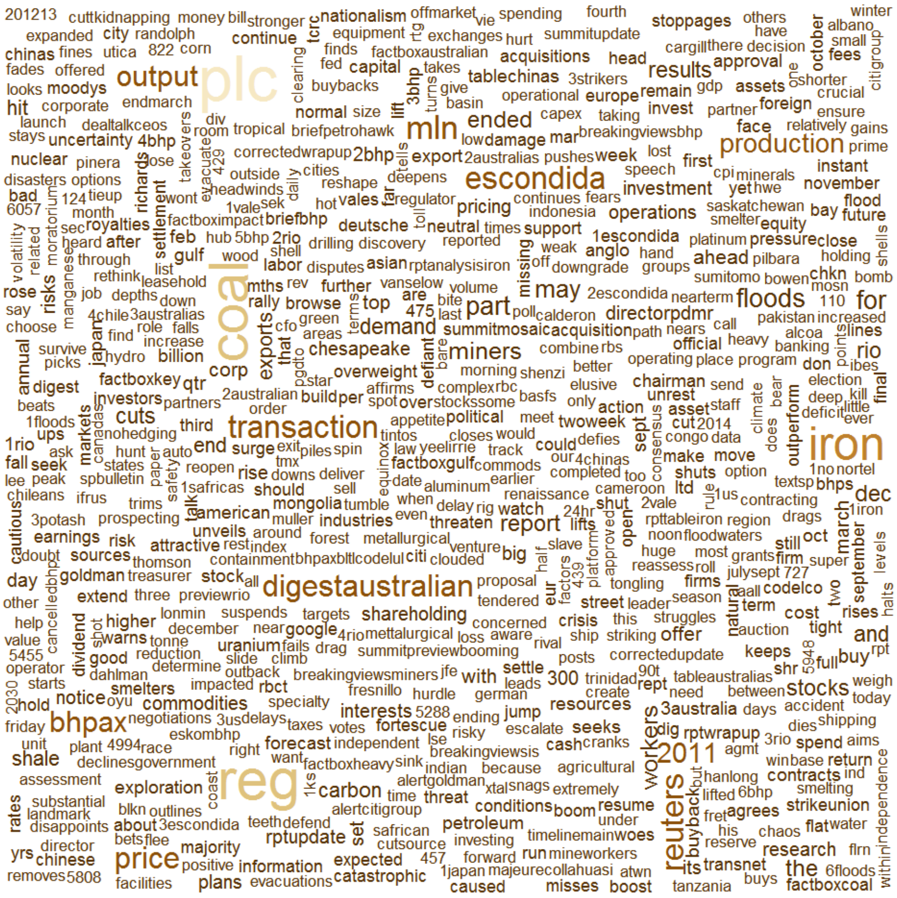

201213 cuttkidnapping money bill stronger nationalism offmarket vie spending fourth stoppages others winter
expanded city randolph continue equipment exchanges hurt summitupdate have albano
chinas fines utica 822 corn finds factboxaustralian acquisitions head cargill there decision october small fees
fades offered output plc fed capital takes tablechinas 3strikers results approval citigroup
looks moodys buybacks 3bhp give europe remain gdp assets shorter crucial
hit corporate endmarch clearing normal size lift turns basin operational invest partner foreign ensure
launch dealtalkceos div briefpetrohawk mln ended capex taking relatively gains
stays uncertainty 4bhp room tropical lowdamage mar breakingviewsbhp face production prime
nuclear pinera close takeovers correctedwrapup 2bhp tells export 2australias pushes week lost first instant
disasters options evacuate 429 outside cities deepens escondida speech cpi minerals
bad 124 tieup richards headwinds reshape daily investment yet hwe november
6057 month wont 1vale sek hot vales regulator pricing continues fears indonesia flood
royalties settlement factboximpact briefbhp deutsche toll neutral times operations saskatchewan bay future
sec heard after feb hub 5bhp 2rio drilling discovery reported support smelter equity
volatility related moratorium through gulf wood shell labor disputes asian rptanalysisiron missing weak 1escondida platinum pressure close
rethink list leasehold mths rev vanselow off downgrade anglo hand groups ahead pilbara holding shells
rose job depths coal rally browse further are volume may 2escondida sumitomo bowen chkn bomb
say risks manganese down 4chile 3australias exports top 475 bite last part nearterm floods mosn 110 for
choose role falls cfo green terms poll calderon directorpdmr call pakistan increased
annual find increase areas demand summitmosaicacquisition path nears alcoa lines
survive japan hydro billion that chesapeake miners combine rbs official heavy banking points rio
picks pgdto bare morning better operating place program don ibes
digest factboxkey qtr corp star overweight defiant shenzi elusive chairman send bear election
beats investors partners 2australian build per complexrbc affirms basfs only unrest deep kill final
1floods markets canada cuts order spot overstockssome appetite political action asset staff climate deficit little
1rio ups nohedging third transaction tintos closes would meet twoweek cut 2014 outperform ever
ask hunt auto end surge exit piles spin law yeelirrie track could defies consensus sept congo data iron
fall seek states paper safety reopen rise tmx factboxgulf commods our 4chinas make move does
lee peak spbulletin 1safricas downs deliver equinox date earlier completed too shuts option 1no nortel
chileans ifrus talk should sell aluminum renaissance cameroon ltd rule 1us textsp bhps
cautious trims american mongolia tumble when delay rig watch 24hr shut 2vale contracting dec
3potash prospecting muller industries even threaten report lifts rpttableiron region drags 1iron
earnings risk unveils around forest metallurgical venture slave open noon floodwaters march
doubt sources attractive rest index approved still oct
thomson containmentbhpaxbltcodelulciti clouded big half factors platform huge most grants firm super
day goldman treasurer stock all digestaustralian 439 reassess roll julysept 727 september
other cancelledbhp extend three previewrio proposal eur tongling firms natural aall codelco two levels
help dividend shot higher lonmin suspends targets shareholding tendered street leader season term
value warns tonne december near google concerned crisis this struggles auction cost rises
5455 good reduction uranium 4rio mettalurgical loss aware ship striking offer tight halts
operator dahlman determine fails drag summitpreviewbooming rival posts correctedupdate keeps and rpt
2030 starts smelters slide climb outback breakingviewsminers jfe with settle 90t 5948 full buy
hold notice oyu commodities rbct fresnillo hurdle leads 300 trinidad tableaustralias shr weigh
friday bhpax impacted specialty interests 5288 german create rept need between stocks today
unit negotiations 3us delays taxes votes fortescue ending jump resources 3australia days accident
shale plant 4994 race eskombhp right forecast independent risky escalate seeks workers dig rptwrapup dies shipping
assessment declines government want factboxheavy sink lse breakingviewsis cash cranks 2011 3rio agmt spend aims
rates substantial exploration coast reg 1ks carbon alertgoldman indian because agricultural reuters hanlong winbase return
landmark blkn outlines time xtal snags extremely lifted 6bhp contracts ind
disappoints about 3escondida teeth defend alertcitigroup threat conditions boom resume back fret agrees smelting
yrs director bets flee rptupdate set safrican petroleum timelinemain woes under buyback strikeunion independence
chinese majority cutsource investing forward run mineworkers his chaos flat water
removes 5808 price positive information expected 457 1japan majeurecollahuasi atwn its transnet reserve research flrn
facilities plans evacuations catastrophic caused misses boost tanzania buys the 6floods within
factboxcoal

Figure 2: TRNA-Sentiment Scores Generated for BHP Billiton in Jan-2011

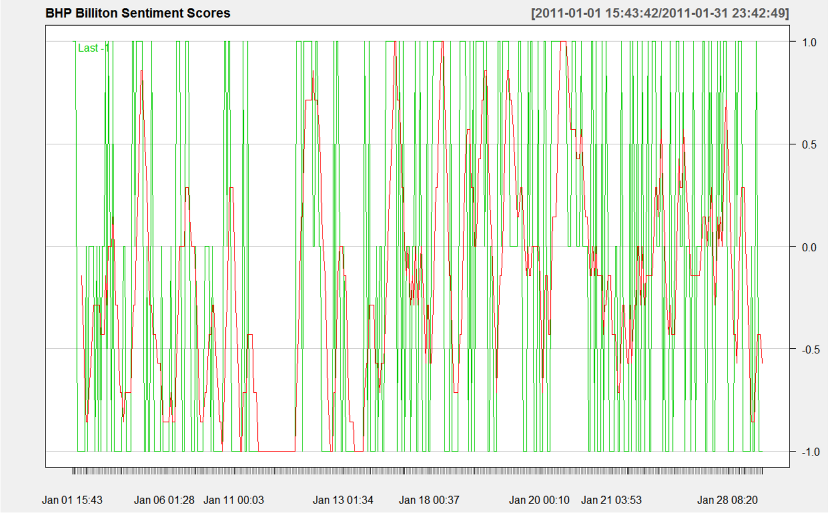


Figure 3: Sentiment Score Distribution for DJIA Stocks in 2008

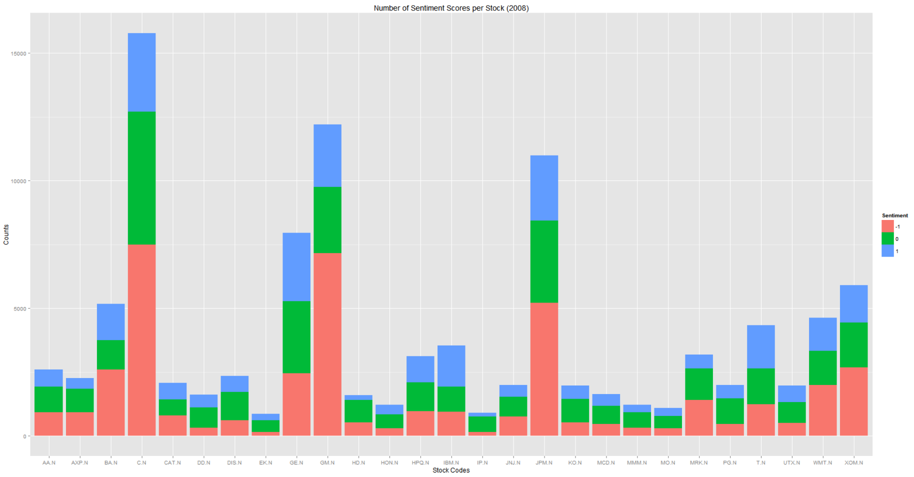
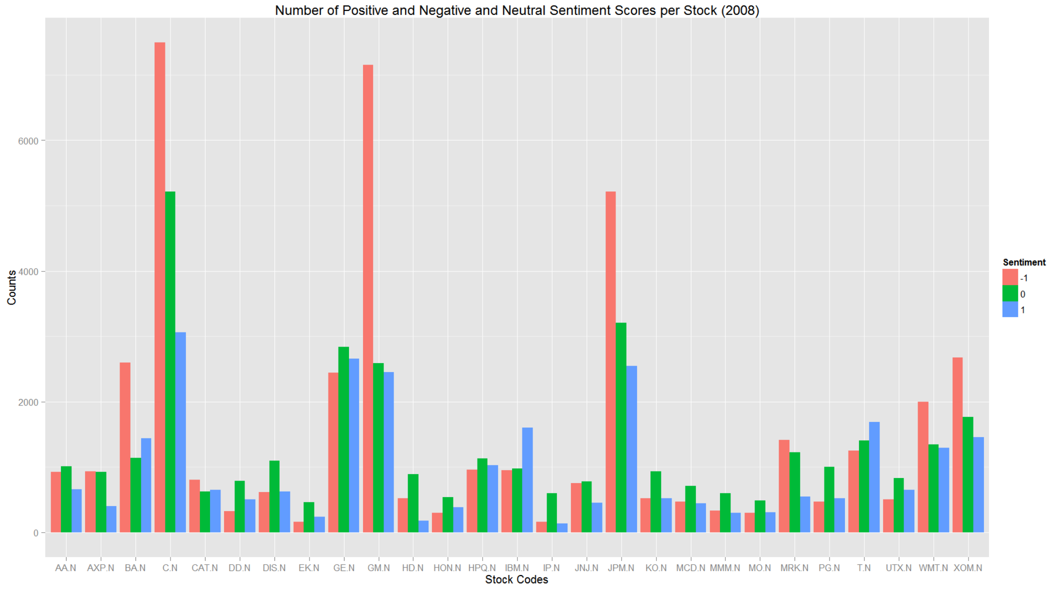


Figure 4: Positive, Negative and Neutral Sentiment Score Distribution for DJIA Stocks in 2008



the effect of news sentiment on DJIA stocks at a daily frequency. We construct daily sentiment index score time series for the empirical exercise from the high frequency scores reported by TRNA.

The empirical analysis in this study analyses the effect of news sentiment on stock prices of the DJIA constituents by considering the daily DJIA market sentiment as an additional risk factor to explain stock returns. We construct daily sentiment scores for DJIA market by accumulating high frequency sentiment scores of the DJIA's constituents obtained from TRNA dataset. We use data from January 2007 to October 2012 to study the senstitivity of the daily stock returns to the daily market sentiment scores. The daily stock prices for all the DJIA traded stocks are obtained from Thomson Tick History database for the same time period.

The TRNA provides high frequency sentiment scores calculated for each news item reported for various stocks and commodities. These TRNA scores for the stocks traded in DJIA can be aggregated to obtain a daily market sentiment score series for the DJIA stock index components. A news item s_t received at time t for a stock is classified as a positive $(+1)$, negative (-1) or neutral (0). $I^+_{s_t}$is a positive classifier (1) for a news item s_t and $I^-_{s_t}$ is the negative (-1) classifier for a news item s_t. TRNA reported sentiment scores have a probability level associated with them, $prob^+_{s_t}$, $prob^-_{s_t}$, $prob^0_{s_t}$ for positive, negative and neutral sentiments, which is reported by TRNA in the Sentiment field. Based on the probability of occurrence, denoted by P_{s_t} for a news item s_t, all the daily sentiments can be combined to obtain a daily sentiment indicator. We use the following formula to obtain the combined score.

$$S = \frac{\sum_{q=t-1}^{t-Q} I^{+}_{s_q} P_{s_q} - \sum_{q=t-1}^{t-Q} I^{-}_{s_q} P_{s_q}}{n_{prob^{+}_{s_q}} + n_{prob^{-}_{s_q}} + n_{prob^{0}_{s_q}}} \tag{1}$$

The time period considered are $t - Q, \ldots, t - 1$ which covers all the news stories (and respective scores) for a 24 hour period.

2.2. Our sample characteristics and preliminary analysis

Table-1 lists the various stocks traded in DJIA along with their RIC (Reuters Instrument Code) and time period. We use the TRNA sentiment scores related to these stocks to obtain the aggregate daily sentiment for the market. The aggregated daily sentiment score S represents the combined score of the sentiment scores reported for the stocks on a particular date. We construct daily sentiment scores for DJIA market by accumulating high frequency sentiment scores of the DJIA's constituents obtained from TRNA dataset. We use data from January 2006 to October 2012 to study the sensitivity of the daily stock returns to the daily market sentiment scores. The daily stock prices for all the DJIA traded stocks are obtained from Thomson Tick History database for the same time period which are provided by SIRCA (The Securities Industry Research Centre of the Asia Pacific).

The stocks with insufficient data are removed from the analysis and the stocks prices for EK.N and EKDKQ.PK are combined together to get a uniform timeseries.

The summary statistics in Table 3 show that our sample of Sentiment scores for the full sample is preominantly negative with a mean of -0.034532. The minimum score is -0.52787 and the maximum score is 0.28564. It appears that negative news is given more prominence than positive news on our scale running from +1 to -1. The Hurst exponent for the Sentiment score with a value of 0.925828 suggests that there is long term memory or persistence in the scores, which makes intuitive sense, given that items of news may take several days to unfold, as greater scrutiny of a story leads to more disclosure of information, and the event, classified as being positive or negative, will tend to occupy the media for several days. This is consistent with trending behaviour. The Hurst exponent for the DJIA is 0.557638 which suggests that the DJIA shows much less tendency to display memory and, as might be expected, behaves more like a random walk. The significant Jarque-Bera test statistics for both series suggest that both are non-Gaussian.

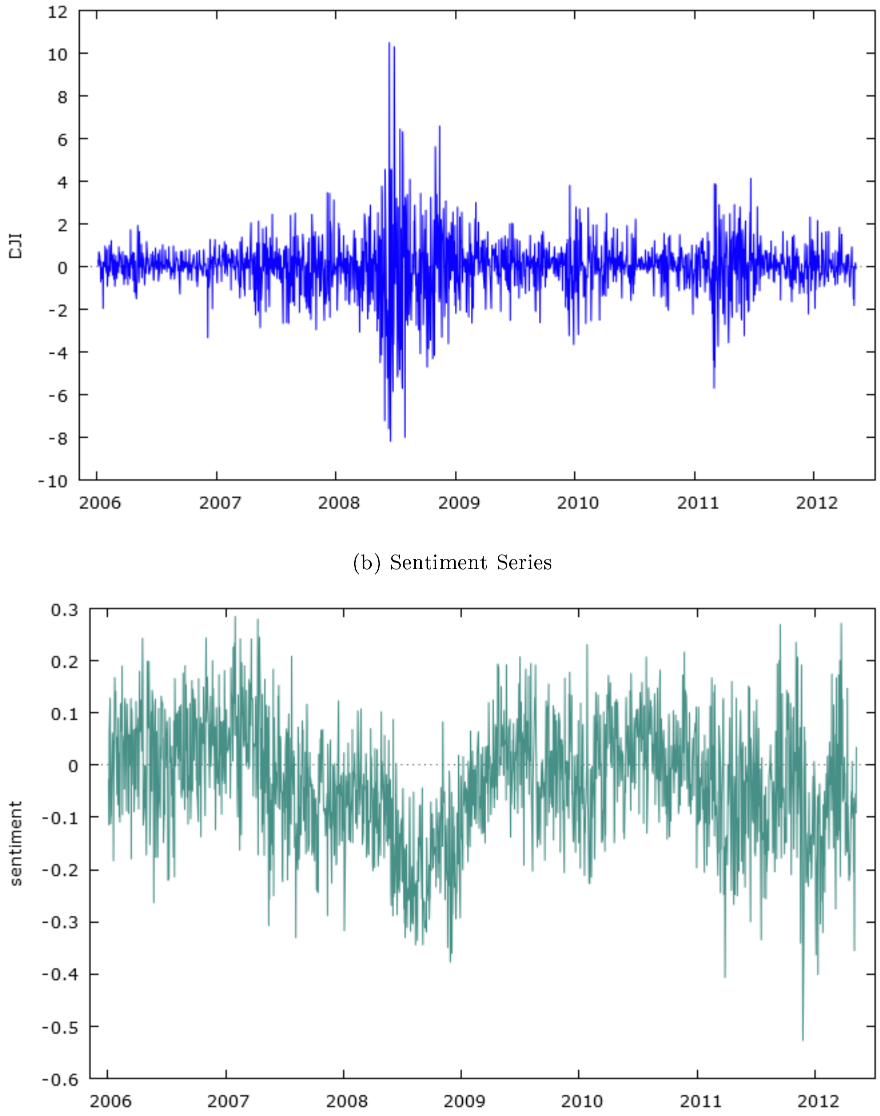
For the one and half years of the sample we have used to capture the height of the GFC in the US market the mean Sentiment score is -0.108907. However, the standard deviation of the sentiment score is 0.106229 which is lower than the value of 0.116762 for the full sample. The Hurst exponent is higher at 0.930411, showing even stronger ternding behaviour, whilst the Jarque-Bera is insignificant, suggesting the distribution cannot be distinhuished from a normal one.

Table 1: DJIA Stocks with Thomson Tick History RIC Codes

RIC Code	Stocks	First Date	Last Date
.DJI	Dow Jones INDU AVERAGE	1-Jan-96	17-Mar-13
AA.N	ALCOA INC	2-Jan-96	18-Mar-13
GE.N	GENERAL ELEC CO	2-Jan-96	18-Mar-13
JNJ.N	JOHNSON&JOHNSON	2-Jan-96	18-Mar-13
MSFT.OQ	MICROSOFT CP	20-Jul-02	18-Mar-13
AXP.N	AMER EXPRESS CO	2-Jan-96	18-Mar-13
GM.N	GENERAL MOTORS	3-Jan-96	18-Mar-13
GMGMQ.PK	GENERAL MOTORS	2-Jun-09	15-Aug-09
JPM.N	JPMORGAN CHASE	1-Jan-96	18-Mar-13
PG.N	PROCTER & GAMBLE	2-Jan-96	18-Mar-13
BA.N	BOEING CO	2-Jan-96	18-Mar-13
HD.N	HOME DEPOT INC	2-Jan-96	18-Mar-13
KO.N	COCA-COLA CO	2-Jan-96	18-Mar-13
SBC.N	SBC COMMS	2-Jan-96	31-Dec-05
T.N	AT&T	3-Jan-96	18-Mar-13
C.N	CITIGROUP	2-Jan-96	18-Mar-13
HON.N	HONEYWELL INTL	2-Jan-96	18-Mar-13
XOM.N	EXXON MOBIL	1-Dec-99	18-Mar-13
MCDw.N	MCDONLDS CORP	6-Oct-06	4-Nov-06
MCD.N	MCDONALD'S CORP	1-Jan-96	18-Mar-13
EK.N	EASTMAN KODAK	1-Jan-96	18-Feb-12
EKDKQ.PK	EASTMAN KODAK	19-Jan-12	18-Mar-13
IP.N	INTNL PAPER CO	2-Jan-96	18-Mar-13
CAT.N	CATERPILLAR INC	2-Jan-96	18-Mar-13
HPQ.N	HEWLETT-PACKARD	4-May-02	18-Mar-13
MMM_w.N	3M COMPANY WI	18-Sep-03	27-Oct-03
MMM.N	MINNESOTA MINIhNG	1-Jan-96	18-Mar-13
UTX.N	UNITED TECH CP	2-Jan-96	18-Mar-13
DD.N	DU PONT CO	2-Jan-96	18-Mar-13
IBM.N	INTL BUS MACHINE	2-Jan-96	18-Mar-13
MO.N	ALTRIA GROUP	2-Jan-96	18-Mar-13
WMT.N	WAL-MART STORES	2-Jan-96	18-Mar-13
DIS.N	WALT DISNEY CO	2-Jan-96	18-Mar-13
INTC.OQ	INTEL CORP	20-Jul-02	18-Mar-13
MRK.N	MERCK & CO	2-Jan-96	18-Mar-13

Table 2: Basic Series Plots: DJIA and Sentiment Scores

(a) DJIA Returns



(b) Sentiment Series

Table 3: Summary statistics, DJIA returns and Sentiment Scores,

Jan 4th 2006 to 31st October 2012			GFC period July 1st 2007-Dec 31st 2008	
	DJIA return (%)	Sentiment Score	DJIA return (%)	Sentiment Score
Min	-8.2005	-0.52787	-8.20051	-0.377528
Median	0.053410	-0.031140	-0.0406688	-0.0996389
Mean	0.013971	-0.034532	-0.100369	-0.108907
Maximum	10.5083	0.28564	10.5083	0.209594
St. Deviation	1.3640	0.116762	2.15087	0.106229
Hurst Exponent	0.557638	0.925828	0.531313	0.930411
Jarque-Bera test	5320.84 (0.00)	18.2197 (0.00)	261.558(0.00)	3.30312(0.19)

Table 4: OLS Regression Results

Model 2: OLS						Model 3: OLS					
Dependent variable: DJI						Dependent variable: DJI					
HAC standard errors						HAC standard errors					
	Coefficient	Std. Error	t-ratio	p-value			Coefficient	Std. Error	t-ratio	p-value	
const	0.101824	0.0257163	3.9595	0.00008	***	const	0.561535	0.107261	5.2352	<0.00001	
sentiment	2.54408	0.311026	8.1796	<0.00001	***	sentiment	6.07768	0.86217	7.0493	<0.00001	***
Mean dependent var	0.013971		S.D. dependent var	1.364036		Mean dependent var	-0.100369		S.D. dependent var	2.150874	
Sum squared resid	2933.248		S.E. of regression	1.331701		Sum squared resid	1738.494		S.E. of regression	2.054178	
R-squared	0.047426		Adjusted R-squared	0.04685		R-squared	0.090101		Adjusted R-squared	0.087892	
F(1, 1654)	66.90636		P-value(F)	5.61E-16		F(1, 412)	49.69238		P-value(F)	7.62E-12	
Log-likelihood	-2823.134		Akaike criterion	5650.268		Log-likelihood	-884.4667		Akaike criterion	1772.933	
Schwarz criterion	5661.093		Hannan-Quinn	5654.281		Schwarz criterion	1780.985		Hannan-Quinn	1776.118	
rho	-0.123209		Durbin-Watson	2.246362		rho	-0.146272		Durbin-Watson	2.288358	

The results of a simple regression of the DJIA return on the Sentiment score for the two periods is shown in Table 4. It can be seen there that the coefficient on Sentiment is highly significant in both periods with a value of 2.54408 for the whole period which rises to 6.007768 during the GFC. The adjusted R Square is higher during the period of the GFC with a value of 0.087892 as compared to 0.04685 for the whole period. The F statistics are highly significant for both periods.

Regression analysis is based on Gaussian assumptions which may not be appropriate. In the next section we will introduce some entropy based metrics before proceeding to present the results of their application in section 4.

3. Entropy-based measures

One attractive feature of the entropy-based set of measures is that they are distribution free. The concept of entropy has its origins in physics in the 19th century and is related to the second law of thermodynamics which states that the entropy of a system cannot decrease other way than by increasing the entropy of another system. This means that the entropy of a system in isolation can only increase or remain constant over time. If the stock market is regarded as a system, then it is not an isolated system: there is a constant transfer of information between the stock market and the real economy. Thus, when information arrives from (leaves to) the real economy, then we can expect to see an increase (decrease) in the entropy of the stock market, corresponding to situations of increased (decreased) randomness.

The most frequent applications of entropy are captured in one of the two main approaches; either as Shannon Entropy – in the discrete case – or as Differential Entropy – in the continuous time case. Shannon Entropy quantifies the expected value of information contained in a realization of a discrete random variable. Shannon entropy can also be used as a measure of uncertainty, or unpredictability: for a uniform discrete distribution, when all the values of the distribution have the same probability, Shannon Entropy reaches its maximum. The minimum value of Shannon Entropy corresponds to perfect predictability, while higher values of Shannon Entropy correspond to lower degrees of predictability. The entropy is a more general measure of uncertainty than the variance or the standard deviation, since the entropy depends on more characteristics of a distribution than does the variance and may be related to the higher moments of a distribution.

A second feature of entropy as a metrc is that whilst both the entropy and the variance reflect the degree of concentration for a particular distribution, their metric is different. This is because the variance measures the concentration around the mean, whilst the entropy measures the diffuseness of the density irrespective of the location parameter. In information theory, entropy is a measure of the uncertainty associated with a random variable. The concept as developed by Shannon (1948) in his use of entropy, was to quantify the expected value of the information contained in a message, which can be measured in units such as bits. In this context, a 'message' means a specific realization of the random variable. The USA National Science Foundation workshop (2003, p. 4) pointed out that the; "Information Technology revolution that has affected Society and the world so fundamentally over the last few decades is squarely based on computation and communication, the roots of which are respectively Computer Science (CS) and Information Theory (IT)". Shannon.(1948) provided the foundation for information theory. In the late 1960s and early 1970s, there were tremendous interdisciplinary research activities from IT and CS, exemplified by the work of Kolmogorov, Chaitin, and Solomonoff, with the aim of establishing algorithmic information theory. Motivated by approaching the Kolmogorov complexity algorithmically, A. Lempel (a computer scientist), and J. Ziv (an information theorist) worked together in later 1970s to develop compression algorithms that are now widely referred to as Lempel-Ziv algorithms. Today, these are the standard approach for lossless text compression. They have broad application in computers, modems, and communication networks. Shannon's entropy represents an absolute limit on the best possible lossless compression of any communication, under certain constraints. It treats messages to be encoded as a sequence of independent and identically-distributed random variables. Shannon's source coding theorem shows that, in the limit, the average length of the shortest possible representation to encode the messages in a given alphabet is their entropy divided by the logarithm of the number of symbols in the target alphabet. For a random variable X with $n\,outcomes$, $\{x_i : i = 1, ...n\}$ the Shannon entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i) \tag{2}$$

where $p(x_i)$is the probability mass function of outcome x_i. Usually logs to the base 2 are used when we are dealing with bits of information. We can also define the joint entropy of two random variables as follows:

$$H\left[X,Y\right] = -\sum_{x\in\mathcal{X}}\sum_{y\in\gamma} Pr(x,y) log_2(Pr(x,y)) \tag{3}$$

The joint entropy is a measure of the uncertainty associated with a joint distribution. Similarly, the conditional entropy can be defned as:

$$H\left[X \mid Y\right] = -\sum_{x\in\mathcal{X}}\sum_{y\in\gamma} Pr(x,y) log_2 Pr(x \mid y) \tag{4}$$

Where the conditional entropy measures the uncertainty associated with a conditional probability. Clearly, a generalised measure of uncertainty has lots of important implications across a wide number of disciplines. In the view of Jaynes (1957), thermodynamic entropy should be seen as an application of Shannon's information theory. Jaynes (2003) gathers various threads of modern thinking about Bayesian probability and statistical inference, develops the notion of probability theory as extended logic and contrasts the advantages of Bayesian techniques with the results of other approaches.

Golan (2002) provides a survey of information-theoretic methods in econometrics and examines the connecting theme among these methods, whilst providing a more detailed summary and synthesis of the sub-class of methods that treat the observed sample moments as stochastic. Granger, Massoumi and Racine (2004) applied estimators based on this approach as a dependence metric for nonlinear processes. Pincus (2008) demonstrates the utility of approximate entropy (ApEn), a model-independent measure of sequential irregularity, via several distinct applications, both to empirical data and in the context of models. He also considers cross-ApEn, a related two-variable measure of asynchrony that provides a more robust and ubiquitous measure of bivariate correspondence than does correlation, and the resultant implications for diversification strategies and portfolio optimisation. A theme further explored by Bera and Park (2008). Sims (2005) discusses information theoretic approaches that have been taken in the existing economics literature to applying Shannon capacity to economic modelling, whilst both critiquing existing models and suggesting promising directions for further progress.

Usually, the variance is regarded as being the central measure in the risk and uncertainty analysis in financial markets. However, the entropy measure can be used as an alternative measure of dispersion, and some authors consider that the variance should be interpreted as a measure of uncertainty with some precaution [see, e.g. Maasoumi (1993) and Soofi (1997)]. Ebrahimi, Maasoumi and

Soofi (1999) examine the role of the variance and entropy in ordering distributions and random prospects, and conclude that there is no general relationship between these measures in terms of ordering distributions. They found that, under certain conditions, the ordering of the variance and entropy is similar for transformations of continuous variables, and show that the entropy depends on many more parameters of a distribution than the variance. Indeed, a Legendre series expansion shows that the entropy is related to higher-order moments of a distribution and thus, unlike the variance, could offer a better characterization of $p_X(x)$ since it uses more information about the probability distribution than the variance [see Ebrahimi *et al.* (1999)].

Maasoumi and Racine (2002) argue that when the empirical probability distribution is not perfectly known, then entropy constitutes an alternative measure for assessing uncertainty, predictability and also goodness-of-fit. It has been suggested that entropy represents the disorder and uncertainty of a stock market index or a particular stock return series, since entropy has the ability to capture the complexity of systems without requiring rigid assumptions that may bias the results obtained.

To estimate entropy in this study we used the 'entropy package' available in the R library, as developed by Hausser and Strimmer (2009). We draw on their account to explain how they develop their estimators: to define the Shannon entropy, they consider a categorical random variable with alphabet size p and associated cell probabilities $\theta_1, ..., \theta_p$, with $\theta_k > 0$ and $\sum_k \theta_k = 1$. If it is assumed that p is fixed and known, then in this case Shannon entropy in natural units is given by:

$$H = -\sum_{k=1}^{p} \theta_k log(\theta_k) \tag{5}$$

In practice the underlying probablity mass function is unknown and therefore H and θ_k need to be estimated from observed cell counts from the sample used $y_k \geq 0$. A commonly used estimator of entropy is the maximum likelihood estimator (ML) which is given by:

$$\hat{H}^{ML} = -\sum_{k=1}^{p} \hat{\theta}_k^{ML} log(\hat{\theta}_k^{ML}) \tag{6}$$

This is formed by substituting in the ML frequency estimates

$$\hat{\theta}_k^{ML} = \frac{y_k}{n} \tag{7}$$

into equation (5), with $n = \sum_{k=1}^{p} y_k$ being the total number of counts.

3.0.1. Maximum Likelihood Estimation

The multinomial distribution is used to make the connection between observed counts y_k and frequencies θ_k.

$$Prob(y_1, ..., y_p) = \frac{n!}{\prod_{k=1}^{p} y_k!} \prod_{k=1}^{p} \theta_k^{y_k} \tag{8}$$

Note that $\theta_k > 0$ otherwise the distribution is singular. By contrast there may be zero counts y_k. The ML estimator of $\theta_k$maximizes the right hand side of equation (8) for fixed y_k, leading to the observed frequencies $\hat{\theta}_k^{ML} = \frac{y_k}{n}$ with variances $Var(\hat{\theta}_k^{ML}) = \frac{1}{n}\theta_k(1 - \theta_k)$ and Bias $(\hat{\theta}_k^{ML}) = 0$ as $E(\hat{\theta}_k^{ML}) = \theta_k$

3.0.2. Miller-Madox Estimator

Even though θ_K^{ML} is unbiased, the plug in entropy estimator $\hat{\theta}^{ML}$ is not. First order bias correction leads to:

$$\hat{H}^{MM} = \hat{H}^{ML} + \frac{m > 0 - 1}{2n} \tag{9}$$

where $m > 0$ is the number of cells with $y_k > 0$. This is temed the Miller-Madow estimator, see Miller (1955).

3.0.3. Bayesian Estimators

Bayesian regularization of the cell counts may lead to improvements over ML estimates. The Dirichlet distribution with parameters $a_1, a_2, ..., a_p$ as prior, the resulting posterior distribution is also Dirichlet with mean

$$\hat{\theta}_k^{Bayes} = \frac{y_k + a_k}{n + A}$$

where $A = \sum_{k=1}^{p} a_k$. The flattening constants a_k play the role of pseudo counts (compare with equation (7)), therefore A may be interpreted as the *a priori* sample size.

$$\hat{H}^{Bayes} = -\sum_{k=1}^{p} \hat{\theta}_k^{Bayes} log(\hat{\theta}_k^{Bayes}) \tag{10}$$

3.0.4. Mutual information

One attraction of entropy-based measures is that they can relax the linearity assumption and capture nonlinear associations amongst variables. The starting point is to capture the mutual information between pairs of variables $MI(X, Y)$. The mutual information is the Kullback-Leibler distance from the joint probability density to the product of the marginal probability densities:

$$MI(X, Y) = E_{f(x,y)} \left\{ log \frac{f(x, y)}{f(x)f(y)} \right\} \tag{11}$$

The measure mutual information (MI) is always non-negative, symmetric, and equals zero only if X and Y are independent. In the case of normally distributed variables MI is closely related to the Pearson Correlation coefficient.

Table 5: Common choices for the parameters of the Dirichlet prior in the Bayesian estimators of cell frequencies, and corresponding entropy estimators

a_k	Cell frequency prior	Entropy estimator
0	no prior	maximum likelihood
1/2	Jeffreys prior (Jeffreys, 1946)	Krichevsky and Trofimov (1981)
1	Bayes-Laplace uniform prior	Holste et al (1998)
1/p	Perks prior (Perks, 1946)	Sh$\ddot{u}$rmann and Grassberger (1996)
$\sqrt{n}/p$	minmax prior (Trybula, 1958)	

Source:Hausser and Strimmer (2009)

$$MI(X,Y) = -\frac{1}{2}log(1-\rho^2)$$

The entropy representations is:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{12}$$

This shows that MI can be computed from the joint and marginal entropies of the two variables.

We use these methods to assess the information content of our two series: DJIA returns and Sentiment plus the degree to which one reveals information about the other. The results are presented in the next section.

4. The results from applying entropy metrics to our basic series.

There are a number of different ways of generating entropy statistics. A key issue is the manner in which prior probabilities are set up. Table 4 presents some of the common methods adopted which are based on different choices of priors.

We used our returns series for DJIA and the Sentiment series and estimated their entropy and cross-entropy. The results are shown in Table 5.

The results in Table 5 are intuitively challenging at first glance. They are divided into statistics for the whole period and for a subset capturing the depths of the financial crisis in the US which constitute a sub-sample running from July 2007 until the end of 2008. The entropy statistics for DJIA returns for the whole sample range from 2.30 to 2.38 depending upon the estimation method used, whilst the entropy statistics for the Sentiment series in the same period range from 3.14 to 3.18. This suggests that there is more uncertainty attached to the Sentiment series than to the DJIA series. For the sub-sample chosen to capture the extremes of the financial crisis the entropy statistics for the DJIA returns range from 2.03 to 2.14, suggesting that there is less uncertainty in terms of this metric, based on Shannon entropy, than in the whole sample period. The same applies to the Sentiment series for which the entropy values range from

Table 6: Entropy and MI statistics for DJIA returns and Sentiment returns

	Whole Sample		Financial Crisis July 1st 2007 -Dec 31st 2008	
	DJIA returns	Sentiment series	DJIA returns	Sentiment series
Maximum Likelihood Estimate	2.300919	3.148384	2.030576	2.680489
Miller-Madow Estimator	2.309675	3.158952	2.049956	2.705037
Jeffrey's prior	2.344529	3.168113	2.092471	2.698928
Bayes-Laplace	2.38224	3.185154	2.143233	2.715458
SG	2.303599	3.14958	2.038168	2.682432
Minimax	2.383488	3.185717	2.141674	2.714943
ChaoShen	2.31512	3.155446	2.052992	2.694162
Mutual Information MI				
MI Empirical (ML)	0.2332534		0.3438407	

2.68 to 2.71. The MI criterion tells a different story, suggesting that there is less uncertainty in the relationship between the two series in the overall period than during the period of the financial crisis.

Ebrahimi, Maasoumi and Soofi (1999) explored the role of entropy and the variance as a method of ordering distributions and random prospects. Their conclusion was that there is no general relationship between the two measures in terms of the ordering of distributions. Under certain conditions, the ordering of the variance and entropy is similar for transformations of continuous variables. However, the entropy of a distribution depends on many more parameters of a distribution than the variance. It can be argued that because the entropy is related to higher-order moments of a distribution it better characterizes a distribution since it uses more information about the probability distribution than the variance [see Ebrahimi et al. (1999)].

The entropy measure captures uncertainty about the behaviour of a probability distribution. Our results in Table 5 suggest that there is less uncertainty about the behaviour of the DJIA and the Sentiment series during the financial crisis than over the period as a whole. A moment's consideration suggests why this could be the case. Our preliminary results in Table 3 revealed that the standard deviation of the Sentiment Score was lower during the GFC than for the whole sample period. The Hurst exponent for Sentiment for the GFC period was also slightly higher suggesting slightly stronger trending behaviour. The distribution for Sentiment was also indistinguishable from a Gaussian distribution for this period. This combines to suggest that in this period the Sentiment score was more predictable, which would be consistent with a lower entropy value.

There is much less difference between the entropy scores across the two periods for the DJIA returns which record a lower entropy score in the GFC across all metrics but it usually differs by less than 0.30. This suggests the behaviour of the DJIA was marginally more predictable during the GFC. The Hurst exponent suggests the reverse and is slightly closer to the value of a random walk at 0.531313 during the GFC than its overall value of 0.557638.

Finally, the MI statistics are consistent with the regression analysis, in that the higher value of 0.3438407 during the GFC compared to 0.2332534 is consistent with the higher adjusted R-squared in the regression during the GFC sub-sample. Though it has to be born in mind that the entropymeasure captures higher moments and non-linearities in a manner that simple, linear OLS does not.

5. Conclusion

In this paper we have analysed the relationship between the TRNA news series for the DJIA constituent stocks after having aggregated them into a daily average Sentiment score time series using all the constituent companies in the DJIA. This was then used in an analysis of the relationship between the two daily sets of series, TRNA news sentiment on the one hand and DJIA returns on the other. We analyse the relationship between the two series using entropy based metrics because these are non-parametric and non-linear and should provide a clear indication of the joint information shared by the two sets of series by means of Mutual Information (MI) metrics. The results of both the summary statistics and simple OLS regression analysis and the non-linear, non-parametric entropy statistics are by and large consistent. A startling result is that there is less uncertainy about the Sentiment series during the GFC period, but on reflection this is perhaps, consistent with intuition. The mean and median sentiment scores are negative, emphasing the newsworthiness of 'bad news', and there is less uncertainty about this bad news during the GFC. The analysis suggests that the behaviour of the DJIA return series, consistent with many previous studies, is much closer to a random walk or Markov process, and consistently displays a relative lack of 'memory'.

Acknowledgements

The authors gratefully acknowledge the support of the QUANTVALLEY/FdR: 'Quantitative Management Initiative'. We are also grateful to SIRCA for providing the TRNA data sets.

References

[1] Barber, B. M., and T. Odean (2008) "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors". *Review of Financial Studies, 21*(2), 785–818.

[2] Bera, A.K. and S. Y. Park (2008) "Optimal Portfolio Diversification Using the Maximum Entropy Principle", *Econometric Reviews*, 27:4-6, 484-512

[3] Borovkova, S., and D. Mahakena (2013) *“News, Volatility and Jumps: The Case of Natural Gas Futures”*. Working Paper. Retrieved From : http://ssrn.com/abstract=2334226

[4] Chaitin, G. J. (1969) "On the Simplicity and Speed of Programs for Computing Infinite Sets of Natural Numbers", *Journal of the ACM* 16 (3): 407

[5] Chao,A., and T. J. Shen (2003) “Nonparametric estimation of Shannon’s index of diversity when there are unseen species”, *Environ. Ecol. Stat.*, 10:429–443, 2003.

[6] Ebrahimi, N., E. Maasoumi, E. Soofi (1999) *Journal of Econometrics*, 90, 2, 317-336.

[7] Golan, A. (2002) "Information and Entropy Econometrics – Editor’s view," *Journal of Econometrics,* 107, 1-2 (2002), 1-15.

[8] Golan. A., and E. Maasoumi (2008) “Information Theoretic and Entropy Methods: An Overview”, *Econometric Reviews*, 27:4-6, 317-328.

[9] Granger, C., E. Maasoumi, and J. Racine (2004) “A dependence metric for possibly nonlinear time series”, *Journal of Time Series Analysis* 25(5), 649-669.

[10] Groß-Klußmann, A., and N. Hautsch (2011) “When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions”. *Journal of Empirical Finance, 18*(2), 321-340

[11] Hayfield,T., and J.S. Racine (2008) “Nonparametric Econometrics: The np Package”, Journal of Statistical Software, 27,5, 1-32.

[12] Hausser, J., and K. Strimmer (2012) *Package ’entropy’*, Repository CRAN, http://www.r-project.org/

[13] Hausser, J., and K. Strimmer (2009) “Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks”, *Journal of Machine Learning Research,* 10, 1469-1484.

[14] Holste, D., I. Große, and H. Herzel (1998) “Bayes’ estimators of generalized entropies”, *J. Phys. A: Math. Gen.*, 31:2551–2566.

[15] Huskaj, B. (2009) “A value-at-risk analysis of VIX futures: Long memory, heavy tails, and asymmetry”. Available at SSRN: http://ssrn.com/abstract=1495229.

[16] Huynh, T. D., and D.R. Smith (2013). News Sentiment and Momentum. FIRN Research Paper.

[17] Jaynes, E.T., (1957) "Information Theory and Statistical Mechanics. II," Physical. Review, 108, 2, 171-190.

[18] Jaynes, E. T. (2003) *Probability Theory: The Logic of Science.* Cambridge University Press, ISBN 0-521- 59271-2.

[19] Jeffreys, H., (1946) "An invariant form for the prior probability in estimation problems," *Proc. Royal. Society (Lond.) A,* 186:453–461.

[20] Kolmogorov, A.N. (1965) "Three Approaches to the Quantitative Definition of Information", *Problems Inform. Transmission* 1 (1): 1–7.

[21] Kolmogorov, A.N. and V. A. Uspensky (1987) "Algorithms and randomness", SIAM J. *Theory of Probability and Its Applications*, vol. 32 389-412.

[22] Krichevsky, R.E., and V. K. Trofimov (1981) "The performance of universal encoding". *IEEE Trans. Inf. Theory,* 27:199–207.

[23] Leinweber, D., and J. Sisk (2011) Relating news analytics to stock returns *The Handbook of News Analytics in Finance* (pp. 147-172): John Wiley & Sons, Ltd.

[24] Maasoumi, E. (1993) "A Compendium to Information Theory in Economics and Econometrics", *Econometric Reviews*, , 12, 2, 137-181

[25] Maasoumi, E., and Racine (2002) "Entropy and Predictability of Stock Market Returns," *Journal of Econometrics*, 107(2), 291–312.

[26] Miller, G.A. (1955) "Note on the bias of information estimates", In H. Quastler, editor, *Information Theory in Psychology II-B*, pages 95–100. Free Press, Glencoe, IL.

[27] Mitra, L., Mitra, G., and D. diBartolomeo (2009) Equity portfolio risk (volatility) estimation using market information and sentiment. *Quantitative Finance, 9*(8), 887–895

[28] National Science Foundation, Report of the National Science Foundation Workshop on Information Theory and Computer Science Interface, Workshop, October 17-18, 2003 Chicago, Illinois

[29] Perks W., (1947) "Some observations on inverse probability including a new indifference rule", *J. Inst. Actuaries*, 73:285–334.

[30] Pincus, S. (2008) "Approximate Entropy as an Irregularity Measure for Financial Data", *Econometric Reviews*, 27:4-6, 329-362

[31] Racine, J.S. (2008) “Nonparametric Econometrics: A Primer”, *Foundations and Trends in Econometrics*, 3, 1, 1–88

[32] Shannon, C.E. (1948) “A Mathematical Theory of Communication”, *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.

[33] Schürmann, T., and P. Grassberger (1996) “Entropy estimation of symbol sequences”, *Chaos,* 6:414–427.

[34] Sinha, N. (2011) *Underreaction to news in the US stock market.* Working Paper. Retrieved From: http://ssrn.com/abstract=1572614

[35] Sims, C.A. (2005) “Rational Inattention: A Research Agenda”, Deutsche Bundesbank Discussion Paper Series 1: Economic Studies No 34/2005

[36] Smales, L. A. (2013). *News Sentiment in the Gold Futures Market.* Working Paper*,* Curtin University of Techonology.

[37] Soofi, E.. (1997) “Information Theoretic Regression Methods”. In: *Advances in Econometrics - Applying Maximum Entropy to Econometric Problems*, Fomby, T. and R. Carter Hill eds. Vol. 12. Jai Press Inc., London.

[38] Solomonoff, R. (1960) "A Preliminary Report on a General Theory of Inductive Inference", Report V-131 (Cambridge, Ma.: Zator Co.). revision, Nov., 1960.

[39] Storkenmaier, A., Wagener, M., and C. Weinhardt (2012) “Public information in fragmented markets”. *Financial Markets and Portfolio Management, 26*(2), 179-215.

[40] Tetlock, P.C. (2007) “Giving content to investor sentiment: the role of media in the stock market”. *Journal of Finance 62*, 1139–1167.

[41] Tetlock, P.C. (2010) “Does public financial news resolve asymmetric information?” *Review of Financial Studies 23*, 3520–3557.

[42] Tetlock, P.C., Macskassy, S.A., and M. Saar-Tsechansky (2008) “More than words: quantifying language to measure firms’ fundamentals”. *Journal of Finance 63*, 1427–1467

[43] Thode Jr., H.C. (2002) *Testing for Normality*, Marcel Dekker, New York.

[44] Trybula, S. (1958) “Some problems of simultaneous minimax estimation”, *Ann. Math. Statist.*, 29:245–253.