SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
INFORMATION THEORY AND COMPUTATION
STATISTICS
COMPUTATIONAL SCIENCE

Received
30 August 2013

Accepted
31 January 2014

Published
27 February 2014

Correspondence and requests for materials should be addressed to I.Z. (ilya.zheludev.09@ucl.ac.uk)

When Can Social Media Lead Financial Markets?

Ilya Zheludev, Robert Smith & Tomaso Aste

Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK.

Social media analytics is showing promise for the prediction of financial markets. However, the true value of such data for trading is unclear due to a lack of consensus on which instruments can be predicted and how. Current approaches are based on the evaluation of message *volumes* and are typically assessed via retrospective (*ex-post facto*) evaluation of trading strategy returns. In this paper, we present instead a *sentiment analysis* methodology to quantify and statistically validate which assets could qualify for trading from social media analytics in an *ex-ante* configuration. We use *sentiment analysis* techniques and Information Theory measures to demonstrate that social media message *sentiment* can contain statistically-significant *ex-ante information* on the future prices of the S&P500 index and a limited set of stocks, in excess of what is achievable using solely message *volumes*.

The proliferation of the Internet into every aspect of our lives has undoubtedly improved our ability to access information in real time. The Internet as we recognise it today has evolved substantially over the last thirty years into a source of information on almost any topic. A particular implementation that has seen substantial growth in the 21st century is social media[1], an example of which is Twitter, a micro-blogging and personal-message sharing service started in 2006. The company, which now handles over 500 million users and over 340 million daily messages, is used globally by a broad demographic[2] to publically broadcast, or 'Tweet' 140-character messages on almost any topic. The implications are that for the first time in human history, it is arguably possible to monitor the moods, thoughts and opinions of a large part of the world's population in an aggregated and real-time manner with almost negligible data-collection costs. Social media data have been used to measure and predict real-world phenomena such as brand popularity[3], motion picture box office returns[4] and election outcomes[5]. Of present focus is the prediction of financial markets via the analysis of Tweets[6–9] and other comparable data sources such as Google Trends[10–12], Yahoo! search engine data[13] and Wikipedia articles[14]. Whilst the rationales behind all these analyses are united together by the existence of information inefficiency in financial markets[15,16], there are still inconsistencies in the effectiveness of these potential predictive indicators. Not only are we still far from a unified consensus on the extent to which financial markets may be predicted in this way, but we are still unaware of what the best methodologies are. Furthermore, the exact range of specific financial assets which could be predicted in this manner is unknown, and neither is the extent to which they can be predicted.

There are at least two schools of thought regarding the best methodologies for assessing financial markets with social media. The first centres on the evaluation of the *volumes* of social media message[8,9], search engine queries[10–13], and Wikipedia views & edits[14]. However, such studies do not quantitatively evaluate the *contents* of social media and Internet text-strings – a valuable source of data – and instead consider just their *volumes*. The second methodology centres on attempts to *lead* financial market movements via the quantitative evaluation of the *content* of social media messages[6,7]. Such methodologies, such as the work by Zhang *et al.* which considers up to 1% of all Tweets[7], attempt to anticipate markets ahead of time are via the concurrent quantitative analysis of the *meaning* of internet messages from large groups of individuals in advance of price changes in financial markets. When applied to the analysis of a group's thoughts on a particular topic, an average estimate from many individuals can offer stronger insights than the viewpoints of just the individual[17]. The computational analysis of the moods of social media messages is one way of ascertaining this "collective wisdom" on a given topic. Known as *sentiment analysis*, the tool is a Natural Language Processing and Opinion Mining subtopic[18,19] which can allow for the classification of the polarity of unstructured text strings with regards to emotional scales, e.g. 'calm' vs. 'anxious'. Thus, the analysis of the *sentiments* of messages could allow for a deeper evaluation of social media's powers to *lead* financial markets, over and above what is possible with solely message-*volume* based analyses. However, the extent of the power of *sentiment analysis* methodologies in financial market prediction applications is still unknown. This is what we investigate in our study, by using rigorous and conservative measures for

Table 1 | Twitter Filters used to collect the social media data. We set-up a custom-built Twitter Collection Framework (TCF) to filter in from up to 10% of all messages from Twitter's elevated-access Gardenhose Feed, those that we deemed to be in reference to the financial instruments we consider in this study. Two types of string-filters were used for stocks: either only industry Ticker-IDs; or industry Ticker-IDs AND/OR Company Names. Other filters, such as those for additional currency pairs or stocks were excluded on the principle of insufficient daily Tweet volumes (<24 per day) as determined prior to the study

Instrument	Filter type	Filter
Apple, Inc. CFDs	Ticker ID AND/OR Company Name	$AAPL AND/OR ''Apple''
Apple, Inc. CFDs	Ticker ID	$AAPL
Amazon.com, Inc. CFDs	Ticker ID AND/OR Company Name	$AMZN AND/OR ''Amazon''
Amazon.com, Inc. CFDs	Ticker ID	$AMZN
American Express, Co. CFDs	Ticker ID AND/OR Company Name	$AXP AND/OR ''American Express''
Bank of America, Corp. CFDs	Ticker ID AND/OR Company Name	$BAC AND/OR ''Bank of America''
Bank of America, Corp. CFDs	Ticker ID	$BAC
Cisco Systems, Inc. CFDs	Ticker ID AND/OR Company Name	$CSCO AND/OR ''Cisco''
EURUSD CFDs	Ticker ID	$EURUSD
EURUSD Futures	Ticker ID	$EURUSD
GBPUSD CFDs	Ticker ID	$GBPUSD
GBPUSD Futures	Ticker ID	$GBPUSD
General Electric, Co. CFDs	Ticker ID AND/OR Company Name	$GE AND/OR ''GE'' AND/OR ''General Electric''
General Electric, Co. CFDs	Ticker ID	$GE
Google, Inc. CFDs	Ticker ID AND/OR Company Name	$GOOG AND/OR ''Google''
Google, Inc. CFDs	Ticker ID	$GOOG
The Home Depot, Inc. CFDs	Ticker ID AND/OR Company Name	$HD AND/OR ''Home Depot''
Hewlett Packard, Co. CFDs	Ticker ID AND/OR Company Name	$HPQ AND/OR ''Hewlett-Packard'' AND/OR ''Hewlett Packard''
Hewlett Packard, Co. CFDs	Ticker ID	$HPQ
IBM, Corp. CFDs	Ticker ID AND/OR Company Name	$IBM AND/OR ''IBM''
IBM, Corp. CFDs	Ticker ID	$IBM
Intel Corp. CFDs	Ticker ID AND/OR Company Name	$INTC AND/OR ''Intel''
Intel Corp. CFDs	Ticker ID	$INTC
Johnson & Johnson, Co. CFDs	Ticker ID AND/OR Company Name	$JNJ AND/OR ''Johnson & Johnson'' AND/OR ''Johnson and Johnson''
J.P. Morgan, Inc. CFDs	Ticker ID AND/OR Company Name	$JPM AND/OR ''JPMorgan'' AND/OR ''JP Morgan''
J.P. Morgan, Inc. CFDs	Ticker ID	$JPM
Coca-Cola, Co. CFDs	Ticker ID AND/OR Company Name	$KO AND/OR ''Coca-Cola'' AND/OR ''Coca Cola''
Coca-Cola, Co. CFDs	Ticker ID	$KO
McDonald's, Corp. CFDs	Ticker ID AND/OR Company Name	$MCD AND/OR ''McDonald's'' AND/OR ''McDonalds''
McDonald's, Corp. CFDs	Ticker ID	$MCD
3M, Co. CFDs	Ticker ID AND/OR Company Name	$MMM AND/OR ''3M''
Microsoft, Corp. CFDs	Ticker ID AND/OR Company Name	$MSFT AND/OR ''Microsoft''
Microsoft, Corp. CFDs	Ticker ID	$MSFT
Oracle, Corp. CFDs	Ticker ID & Company Name	$ORCL AND/OR ''Oracle''
Oracle, Corp. CFDs	Ticker ID	$ORCL
FTSE100 Index CFDs	UK Geographical	String-unfiltered UK Tweets
FTSE100 Index Futures	UK Geographical	String-unfiltered UK Tweets
S&P500 Index CFDs	US Geographical	String-unfiltered US Tweets
S&P500 Index Futures	US Geographical	String-unfiltered US Tweets
AT&T, Inc. CFDs	Ticker ID AND/OR Company Name	$T AND/OR ''AT&T''
AT&T, Inc. CFDs	Ticker ID	$T
Wal-Mart, Inc. CFDs	Ticker ID AND/OR Company Name	$WMT AND/OR ''Wal-Mart'' AND/OR ''Wal Mart''
Exxon Mobil, Corp. CFDs	Ticker ID AND/OR Company Name	$XOM AND/OR ''Exxon Mobil''
Exxon Mobil, Corp. CFDs	Ticker ID	$XOM

statistical-significance to analyse up to 10% of all messages from Twitter's network. The methodology presented in this paper is not a trading strategy, nor is it a prediction-indicator generator. It is a necessary and currently-overlooked precursor to compliment the aforementioned studies in this sphere: we seek to determine whether social media *sentiment* data can *lead* financial markets – and to what extent - without using data-mining analysis approaches and without considering market prediction *per se*. Instead, we are concerned only with whether the *sentiment* of social media messages contains useful *information* about future prices of the assets being discussed, without reference to a particular trading strategy.

We apply our methodology to stocks, currencies, and indices to form an overview of the extent to which social media *sentiment* may contain *ex-ante lead-time information* about financial markets, without any possible bias associated with structuring trading strategies. We present an Information Theory metric, which allows us to determine with statistical-significance the extent to which the *sentiment* of social media messages contain *lead-time information* about securities' hourly returns. Specifically, we compare the hourly changes in the *sentiments* of Tweets from the USA and the UK filtered using forty-four specifically-tailored criteria ('Twitter Filters') with the hourly returns of twenty-eight financial instruments ('financial data') collected over a 3-month period (see Table 1 and the *Supplementary Information*). We consider: CFDs for the biggest US stocks using string-filtering; S&P500 index Futures and CFDs using string-unfiltered US-Tweets; FTSE100 index Futures and CFDs using string-unfiltered UK-Tweets; and the GBPUSD and EURUSD currency pairs (both CFDs and Futures). By instituting *time-shifts* of up to 24-hours such that *sentiment* data *leads* the financial data in advance, we show that within the time period that we investigate, the *sentiments* of Twitter messages contain statistically-significant *lead-time information*

on twelve of these financial-instrument/Twitter-Filter combinations. We provide insights into the *leading time-shifts* for each such statistically-significant financial-instrument/Twitter-Filter combination, as shown in Figure 1. Furthermore, by performing identical analysis using just Tweet message *volumes* – rather than their *sentiments* - we demonstrate that the *sentiment* of social media messages is more statistically-significant in *leading* the financial markets than just message *volumes* in all but one case, as shown in Figure 2.

Results

We analyse the performance of intraday *sentiment* data in comparison to intraday financial data over a 3-month period from 11/Dec/12 to 12/Mar/13 (see *Supplementary Information*). Afterwards, by repeating the same experiments using just Twitter message *volumes* (instead of message *sentiments*), we subsequently determine the extent to which Twitter message *volumes* alone can *lead* the financial data. In this manner we show that the *sentiment* of Twitter messages carries greater powers to *lead* the financial data than Twitter message *volumes*.

The social media data corresponding to individual financial instruments is collected via the use of three Twitter Filter types: 1) instrument Ticker-ID filters, e.g. "$CSCO"; 2) combined instrument Ticker ID and/or Company Name filters, e.g. "$CSCO" AND/OR "Cisco"; 3) string-unfiltered Tweets from the USA and alternatively from the UK. The financial data consisted of intraday Futures prices for indices and currency pairs, and intraday Contracts for Difference (CFDs) prices for indices, currency pairs, and stocks. Tweet *sentiments* were derived using SentiStrength[19], a leading[20] research-orientated, fully-transparent English-language *sentiment* classification system specifically tailored to the often grammatically and lexically-incorrect nature of social media vernacular (see *Supplementary Information*). The system has been found to outperform baseline competitors in terms of the accuracy of ranking the *sentiment* of social media vernacular found on MySpace pages[19], and more recently in ranking the *sentiments* of YouTube video comments, Tweets, and online posts on the Runner's World forum[20]. However we note that SentiStrength is not specifically programmed to accurately rank complex complex elements of human speech such as sarcasm and irony. We used SentiStrength's default configuration to produce three *sentiment* scores for each Tweet: positive *sentiment*; negative *sentiment*; and the overall net resultant *sentiment* score (which is calculated by subtracting the negative *sentiment* from the positive *sentiment* for each message). In each case, the performances of these three *sentiment* types were examined independently against the financial data. To achieve this, the *sentiment* data and the corresponding financial data for each Twitter Filter were aggregated by way of mean averaging into discretised non-overlapping consecutive windows of 1-hour in size. Here, the hourly changes in the *sentiment* data ($\Delta_{Sentiment}$) were calculated relative to the previous time-window. Similarly, the hourly changes in the financial prices (Δ_{Price}) were calculated relative to the previous hour to generate an indication of hourly returns. We thus compare the $\Delta_{Sentiment}$ vs. the Δ_{Price} for each financial-instrument/Twitter-Filter combination.

Autocorrelation within the *sentiment* data. We observed autocorrelation in the *sentiment* data produced from each Twitter-Filter, peaking at a lag of 24-hours. We therefore suggest that social media data are autocorrelated at the 24-hour cycle, and therefore this autocorrelation necessitates its removal prior to further analysis. We argue that this condition is necessary in order to avoid the false identification of relationships being gleaned from the dataset which could be driven just by intrinsic autocorrelation. In our study, the autocorrelative processes were removed by applying a 24-hour backward-looking rolling simple moving average (SMA) to the social media data. For each element in the social media time-series, this was determined by calculating the mean of the preceding twenty-three data points and the element in question. However, for the first twenty-three entries in the social media data time-series – for which there are less than twenty-four preceding elements – we calculate the SMA for each such entry based on the mean of the element itself and all available chronologically-preceding elements, up until the first in the time-series. For example, for element 13 of the social-media time-series D: $SMA_{i=13} = \frac{D_{13}+D_{12}+\cdots D_1}{13}$, whilst for element 42 of the social-media time-series D: $SMA_{i=42} = \frac{D_{42}+D_{41}+\cdots D_{19}}{24}$, (see *Supplementary Information*).

Determining if *sentiment* data *leads* financial data. We use concepts from Information Theory to quantify if social media *sentiment* can *lead* the financial data in a statistically-significant manner. Specifically, we consider the Mutual Information[21] between the two time-series of hourly changes in *sentiment* scores and prices at different *time-shifts*. Mutual Information shows the amount of uncertainty in a time-series which can be removed by observing another time-series. Thus, the greater the Mutual Information between time-series 1 and time-series 2, the more we can establish about the nature of time-series 2 by observing time-series 1. The computation of entropy, which is necessary as part of the process for calculating Mutual Information, is based on the probability distribution of the values within the dataset being investigated. In our study we estimate such probability distributions using a histogram. We select bin-size using Sturges' Histogram Rule[22], a well-known method for histogram binning, and verify that we have tested the robustness of our results with respect to changes in bin sizes, finding non-significant differences. For each financial-instrument/Twitter-Filter combination we first determine the Mutual Information available between *sentiment* data and the corresponding financial data at no *time-shift* (when *sentiment* data and financial data are chronologically superimposed). We then institute a *leading time-shift* between the two time-series, such that hourly changes in the *sentiment* data *precede* hourly changes in the price data, and determine the amount of Mutual Information now available compared to the condition where the *time-shift* between the two time-series was zero.

Suppose that the amount of Mutual Information μ between hourly changes in the *sentiment* data and hourly changes in the price data at a *time-shift* of zero hours $L = 0$ is equal to x: $\mu_{L=0} = x$. Now, suppose that the amount of Mutual Information μ between hourly changes in the *sentiment* data and hourly changes in the price data at a *leading time-shift* of $L > 0$ is equal to y: $\mu_{L>0} = y$. We refer to the percentage increase in Mutual Information between the two aforementioned conditions, $\mu_{\%inc}$ from $\mu_{L=0} = x$ to $\mu_{L>0} = y$ as the *information surplus*. If the *information surplus* is positive, i.e. $\mu_{\%inc} > 0$, then hourly changes in the *sentiment* data contain more Mutual Information about securities' hourly returns at a *leading time-shift* of $L > 0$ than at no *time-shift*, $L = 0$. In such scenario hourly changes in the *sentiment* data contain *lead-time information* about hourly returns as they remove more uncertainty, ahead of time, about the financial data time-series than at no *leading time-shift*. Conversely, if the *information surplus* is negative, i.e. $\mu_{\%inc} < 0$, then hourly changes in the *sentiment* data contain less Mutual Information about securities' hourly returns at a *leading time-shift* of $L > 0$ than at no *time-shift*, $L = 0$. In such scenario *sentiment* data does not contain *lead-time information* about hourly returns as they remove less uncertainty, ahead of time, about the financial data time-series than at no *leading time-shift*. We offset the changes in the *sentiment* data ahead of the securities' returns data from 0-hours to 24-hours in 1-hour increments. We then perform the aforementioned Mutual Information comparisons on the hourly changes in the *sentiment* data (for all three *sentiment* types: positive; negative; and net) and the hourly changes in the price data from all forty-four Twitter Filters using the 24-hour autocorrelation-removal condition described earlier. In this manner we determine the *information surplus* for each

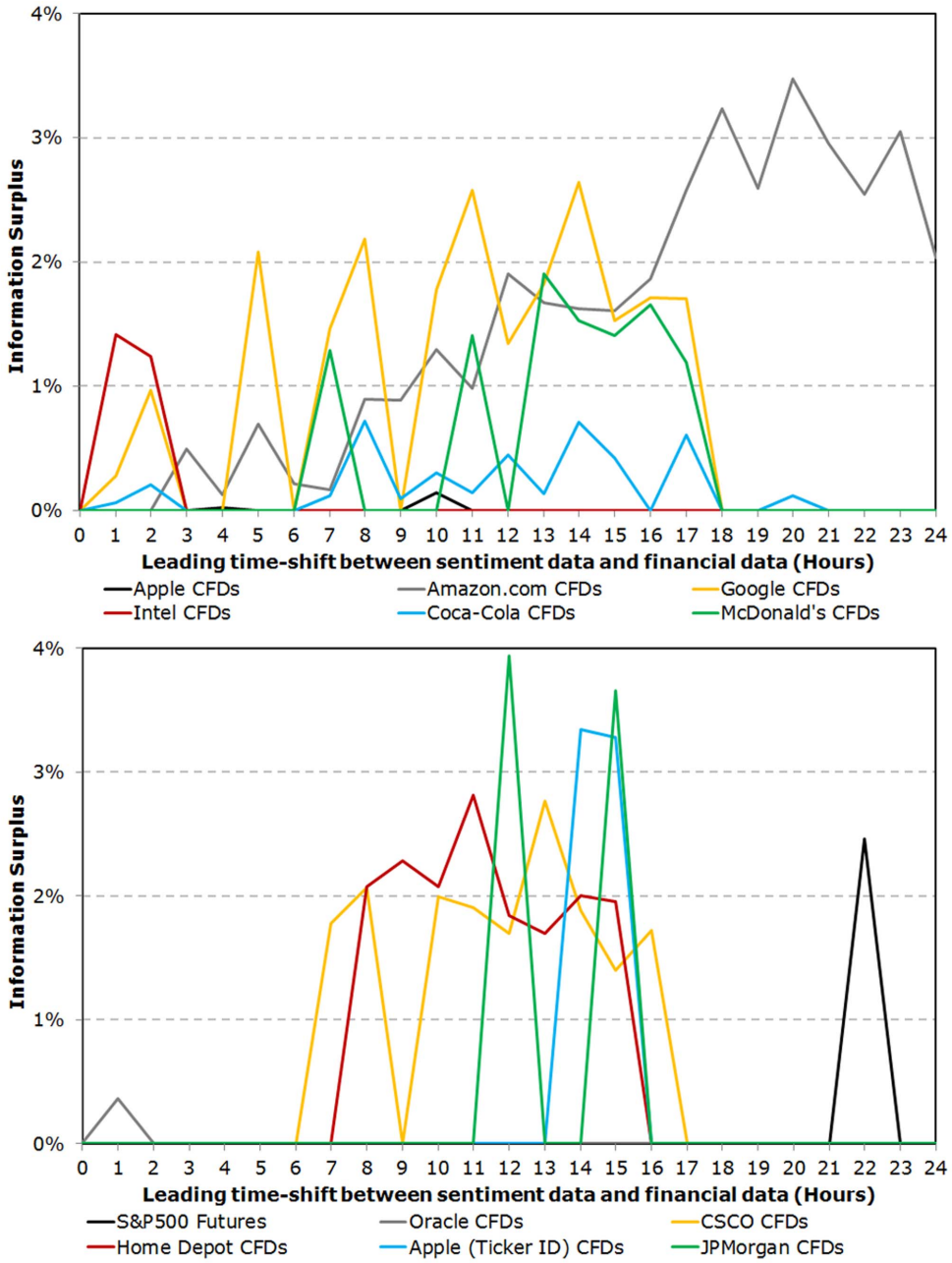

Figure 1 | Examples of when hourly changes in social media *sentiment* contain *lead-time information* securities' hourly returns ahead of time. We refer to the percentage increase in Mutual Information between hourly changes in the social media *sentiment* data and securities' hourly returns at *leading time-shifts*, relative to zero *time-shift*, as the *information surplus*. Here, social media *sentiment* data is offset such that it precedes financial data, and the Mutual Information between the two time-series is compared to that which is available at no *time-shift*. If the *information surplus* is positive, then *sentiment* data contains more Mutual Information about financial data at an exploitable *leading time-shift*, compared with the no-offset configuration. We suggest that in such scenarios, hourly changes in the *sentiment* data contain *lead-time information* about securities' hourly returns as they remove more uncertainty, ahead of time, about the financial data time-series than if the two time-series are not offset. To determine eligibility for social media to *lead* financial data, three further caveats were met: the assets' Twitter Filters attracted a minimum mean message volume of 60 messages per hour from our connection to Twitter's 10% Gardenhose feed; the *information surplus* values were greater when *sentiment* data preceded financial data, than the converse (when financial data preceded *sentiment* data); and finally that the observations were statistically-significant to the 99% confidence interval (relative to *sentiments* generated from randomly permutated data). In this manner, we identify twelve instruments for which hourly changes in the *sentiments* of social media messages contain *lead-time information* about securities' hourly returns ahead of time. In this figure, we show the maximum *information surplus* seen per t*ime-shift*. Of the permitted assets, Apple Inc. was the only company for which such an indication was visible using a Twitter Filter searching solely for an asset's industry Ticker-ID (rather than the company name). Tweets on the remaining individual stocks were obtained by filtering Twitter for Company Names AND/OR their industry Ticker-IDs. Finally, the *sentiments* of string-unfiltered Tweets from the USA were shown to *lead* the returns of S&P500 Futures for one *time-shift*.

Figure 2 | Hourly changes in Tweet message *sentiments lead* financial data more than hourly changes in Tweet message *volumes*. We use Mutual Information to determine the extent to which Twitter messages on financial instruments can *lead* their securities' returns. We perform our analysis on hourly changes in Tweet *sentiments* vs. the hourly returns of forty-four financial instruments, showing that Twitter *sentiment leads* securities' returns in a statistically-significant manner for twelve instruments. We then perform identical analyses on the hourly changes in Twitter message *volumes* vs. the hourly returns and the absolute hourly returns of the same forty-four financial instruments, to echo recent studies which compare social media[8,9] and search engine[10–13] message *volumes* with financial market performance. We demonstrate that the Tweet *sentiments* result in proportionally larger maximum *information surplus* values compared to the maximum *information surplus* values seen from our Tweet *volume* (rather than Tweet *sentiment*) experiments. This is demonstrated in the top chart, where we show the ratios of the maximum *leading statistically-significant information surpluses* seen from our three experiments: hourly changes in Tweet message *sentiments* as evaluated against hourly returns (blue bars); hourly changes in Tweet message *volumes* as evaluated against hourly returns (red bars); and hourly changes in Tweet message *volumes* as evaluated against absolute hourly returns (green bars). Tweet message *sentiments* outperformed Tweet message *volumes* in *leading* securities' hourly returns in a statistically-significant manner for twelve assets. In the bottom chart we demonstrate the ratios of the *number* of observed *instances* of *statistically-significant leading information surpluses* from our three experiments for each asset. We observe that for twelve assets, hourly changes in Tweet message *sentiments* (blue bars) *lead* the securities' hourly returns more often than hourly changes in Tweet message *volumes*, whether these *volumes* are evaluated against hourly returns (red bars) or absolute hourly returns (green bars). In one additional one case (Bank of America, Corp.) hourly changes in Tweet message *volumes led* the security's hourly returns in a statistically-significant manner when Tweet message *sentiments* did not. For all remaining assets from the original forty-four, Tweets do not *lead* securities' returns in a statistically-significant manner.

financial-instrument/Twitter-Filter combination. We can then identify the *leading time-shift*(s), if any, at which hourly changes in the *sentiment* data *lead* the securities' hourly returns. We identify the *sentiment* type (positive; negative; or net) which results in the maximum *information surplus* for each financial-instrument/Twitter-Filter combination. We are thus able to determine the *leading time-shift* for each financial-instrument/Twitter-Filter combination which results in the largest *information surplus*.

Our goal is to determine for which assets *sentiment* data leads financial data in a statistically-significant manner. Thus, three further caveats remain. The first is to ascertain that a sufficiently adequate volume of messages is transferred per hour for each Twitter Filter on each financial instrument to warrant a sufficient statistical sample. Our connection to Twitter permits access to 10% of all Tweets, and therefore with regards to data-density builds on the work of Zhang *et al*. who assessed only 1% of all Tweets[7]. Known as a connection to Twitter's elevated "Gardenhose Feed", it is available free-of-charge for research purposes based on a contractual agreement with Twitter. Considering this limitation in data volume, we propose a minimum viable mean message volume of 1 Tweet per Twitter Filter per minute over the 3-month collection period for our dataset. This would translate to a hypothetical volume of 10 Tweets per minute if access to Twitter's full 100% "Firehose Feed" were available. Based on this message-volume filter, we eliminate twenty-three of the forty-four financial-instrument/Twitter-Filter combinations originally explored. We also exclude the Twitter Filters which reference companies whose names are only two-characters in length, as they attract messages not related to the companies in question. Here, we find that Tweets on the company 3M cannot be filtered accurately since the term "3M" attracts a large volume of messages that have no association with the firm. Similarly, the term "GE" – an often-used trading name of General Electric – attracts large volumes of messages that do not refer this firm either. On the principle of insufficient message volumes, we also reject the following industry Ticker-ID-only Twitter Filters: "$AMZN" (Amazon.com Inc.), "$T" (AT&T Inc.), "$BAC" (Bank of America Corp.), "$KO" (Coca-Cola Co.), "$EURUSD" (currency pair), "$GBPUSD" (currency pair), "$XOM" (Exxon Mobil Corp.), "$GOOG" (Google Inc.), "$HPQ" (Hewlett Packard Co.), "$IBM" (IBM Corp.), "$INTC" (Intel Corp), "$JPM" (J.P. Morgan Inc.), "$MCD" (McDonald's Corp.), "$MSFT" (Microsoft Corp.) and "$ORCL" (Oracle Corp.). Finally, we also reject the following Twitter Filters which use Company Names AND/OR industry Ticker-IDs, also on the principle of insufficient message volume: American Express Co., AT&T Inc., Exxon Mobil Corp., Hewlett Packard Co., and Johnson & Johnson Co.

The second caveat is to ascertain that our *information surplus* methodology is able to identify financial instruments for which the hourly changes in the *sentiment* data carry more *information* about the hourly returns data *before* price changes rather than *after* price changes. In such a manner we could support the notion that *sentiment* data may contain *lead-time information* about financial data rather than merely *reacting* to it. To do this, for each *time-shift* offset of 1-hour to 24-hours between the hourly changes in the *sentiment* data preceding the hourly returns data, we calculate the Mutual Information between the two time-series using the full 24-hour auto-correlation-removal condition, thus identifying the '***per-time-shift*** *leading Mutual Information*' for each financial-instrument/Twitter-Filter combination. We then determine the '***mean*** *trailing Mutual Information*': the mean Mutual Information between the hourly changes in the *sentiment* data and the securities' hourly returns for each financial-instrument/Twitter-Filter combination when offsetting the two time-series so that *sentiment* data *follows* (rather than *leads*) the financial data. We report an example of this in Figure 3. In such a manner we are able to identify instances when for a given *leading time-shift* between the hourly changes in the *sentiment* data and the securities' hourly returns data, social media data is more *leading* than *trailing*. For a given *leading time-shift*, we only admit those financial-instrument/Twitter-Filter combinations for which the *per-time-shift leading Mutual Information* exceeds the *mean trailing Mutual Information*. We then calculate the *information surplus* for each such *leading time-shift* relative to no *time-shift*, and only admit those which result in a positive *information surplus*, as shown by way of example in Figure 4. Conceptually, this filtering mechanism identifies when hourly changes in *sentiment* data carry more *information* about securities' hourly returns ahead of time than at zero *leading time-shift* to show which *time-shifts*, if any, result in *sentiment* data preceding financial data in a manner such that it is more *leading* than *trailing*. A negative *information surplus* would imply that hourly changes in *sentiment* data carry less *information* about securities' hourly returns than at no *time-shift* between the social media and financial data time-series.

The final caveat is to determine the statistical-significance of situations where the hourly changes in the *sentiment* data are shown to be more *leading* than *trailing* for a given *time-shift*. To achieve this, we randomly permutate 10,000 times the hourly changes in *sentiment* data ($\Delta_{Sentiment}$) for each *sentiment* type: positive, negative, or net with respect to the hourly changes in asset price data (Δ_{Price}) and thus calculate the *randomised Mutual Information* at each permutation for a given financial-instrument/Twitter-Filter combination for each *leading time-shift* from 0 hours to 24-hours. We evaluate the *observed Mutual Information* for each sentiment type (positive, negative or net) against the *randomised Mutual Information* for each *sentiment* type independently to avoid a multiple-hypothesis testing configuration. We are thus able to calculate the frequency at which the *observed Mutual Information* between the hourly changes in the *sentiment* data and the securities' hourly returns exceeds the *randomised Mutual Information* over the 10,000 random permutations. We therefore accept those *leading time-shifts* for which the *observed Mutual Information* between the hourly changes in the *sentiment* data and the securities' hourly returns is greater than the *randomised Mutual Information* with a statistically-significant confidence interval of 99%.

Summarising, by satisfying the three aforementioned caveats, we first exclude those financial-instrument/Twitter-Filter combinations which do not attract sufficient hourly Tweet volumes, or which yield incorrect messages due to two-character company names which attract large volumes of irrelevant messages. This leaves nineteen assets on which we then apply the criteria discussed before: to test whether social media *sentiment* is more *leading* than *trailing* when evaluated against financial data at different *time-shifts*; and to test the resultant relationships for statistical-significance. Consequently, we are able to identify statistically-significant *leading time-shifts* for which hourly changes in the *sentiment* data *lead* securities' hourly returns, an example of which is shown in Figure 5.

In such a manner we are able to identify a range of *leading time-shifts* for twelve of the aforementioned remaining nineteen financial-instrument/Twitter-Filter combinations, which demonstrate with statistical-significance the ability for social media *sentiment* to *lead* financial data in certain cases. In Table 2 we offer a summary of these permitted financial instruments, the characteristics of their Twitter Filters, their mean hourly message volume over the 3-month collection period, their largest-observed statistically-significant *leading information surplus* values, the corresponding best-performing *leading time-shifts*, the corresponding optimum *sentiment* type (positive, negative or net), and the number of statistically-significant *leading time-shifts* identified during this investigation. The spectrum of statistically-significant *leading information surplus* values seen for each of these twelve financial-instrument/Twitter-Filter combinations admitted by this *sentiment analysis* experiment is shown in Figure 1. Here, we observe that the number of instances of Tweet *sentiment* leading financial data is heterogeneous across these twelve assets – for example we detect only one instance of string-unfiltered Tweets from the USA *leading* S&P500 index Futures, but twenty-two counts of Tweets filtered by "Amazon" AND/OR "$AMZN" *leading* Amazon.com, Inc. CFDs.

In order to determine if hourly changes in *sentiment* data carry more *information* than hourly changes in just Tweet *volumes*, we repeat our experiments using just Tweet message *volumes* rather than Tweet *sentiments*. We evaluate $\Delta_{Tweet\ volume}$ against Δ_{Price} (the hourly

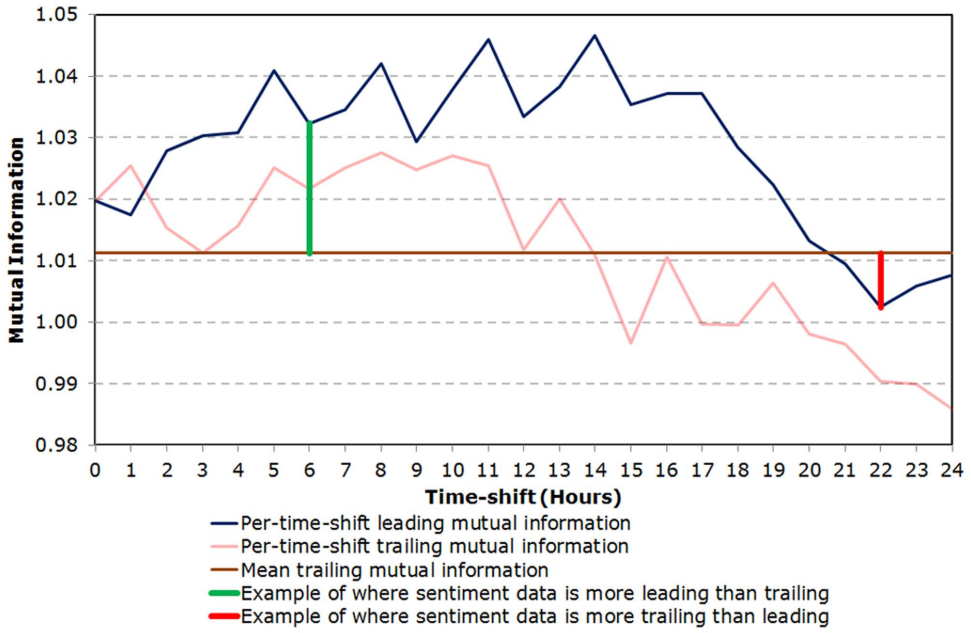

Figure 3 | Determining if *sentiment* data is more *leading* than *trailing*. By way of example, we demonstrate the Mutual Information between hourly changes in *sentiments* and financial data for the Twitter Filter: "$GOOG" AND/OR "Google" compared with the hourly returns of Google CFDs. For this example, we only consider the negative *sentiments* as calculated by SentiStrength, a leading[20] research-orientated *sentiment* classification tool tailored for the lexically and grammatically-incorrect nature of social media text. The data is presented for *time-shifts* between 0 and 24-hours both in a *leading* configuration (such that hourly changes in the *sentiment* data *lead* the security's hourly returns) and in a *trailing* configuration (such that security's hourly returns *lead* the hourly changes in the *sentiment* data). We only admit those *time-shifts* for which the *per-time-shift leading Mutual Information* exceeds the *mean trailing Mutual Information*, as indicated by the vertical green bar, and reject those *time-shifts* for which *per-time-shift leading Mutual Information* is less than the *mean trailing Mutual Information*, as indicated by the vertical red bar.

returns) for each financial-instrument/Twitter-Filter combination to evaluate the extent to which hourly changes in Tweet *volumes* can *lead* the securities' hourly returns using our methodology as an echo of past studies which compare social media[8,9] and search engine[10–13] message *volumes* with financial market performance. We then also repeat this experiment to consider $\Delta_{Tweet\ volume}$ against $|\Delta_{Price}|$ (the absolute hourly returns) to further explore the ability of hourly changes in Tweet *volumes* to *lead* securities' hourly returns. We

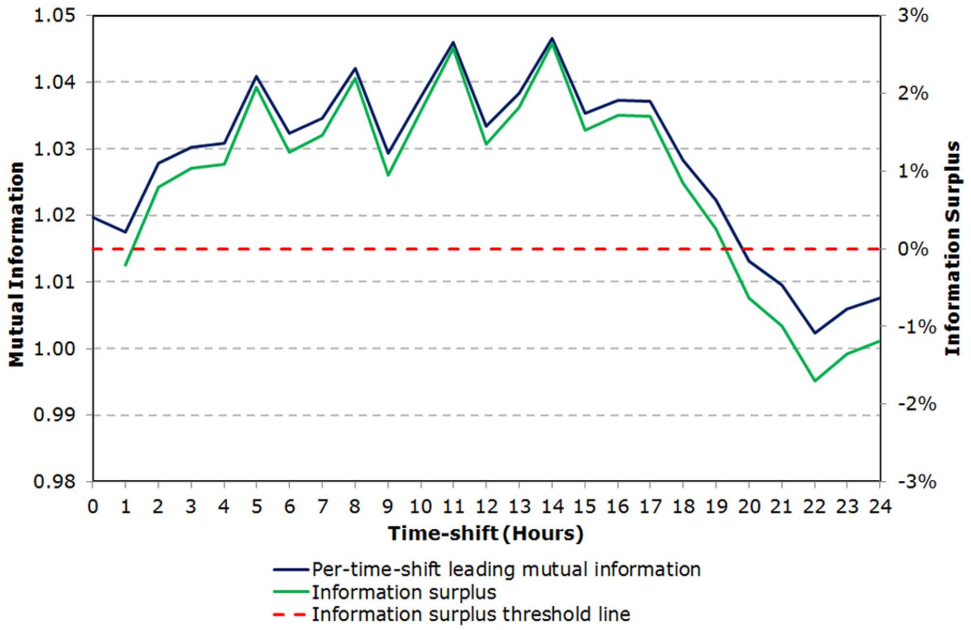

Figure 4 | Determining if *sentiment* data can *lead* financial data. We use the term *information surplus* to denote situations when hourly changes in the *sentiment* data carry more *information* about securities' hourly returns ahead of time than at no *leading time-shift*. By way of example, we demonstrate the *information surplus* between hourly changes in the *sentiment* data for the Twitter Filter: "$GOOG" AND/OR "Google" and the hourly returns of Google, Inc. CFDs. For the *sentiment* data to be considered *leading*, it must demonstrate positive *information surplus* at *time-shifts* where *sentiment data* is offset to *lead* financial data. As in the example above, we admit those *leading time-shifts* for which the *information surplus* curve is above the *information surplus threshold line* of zero.

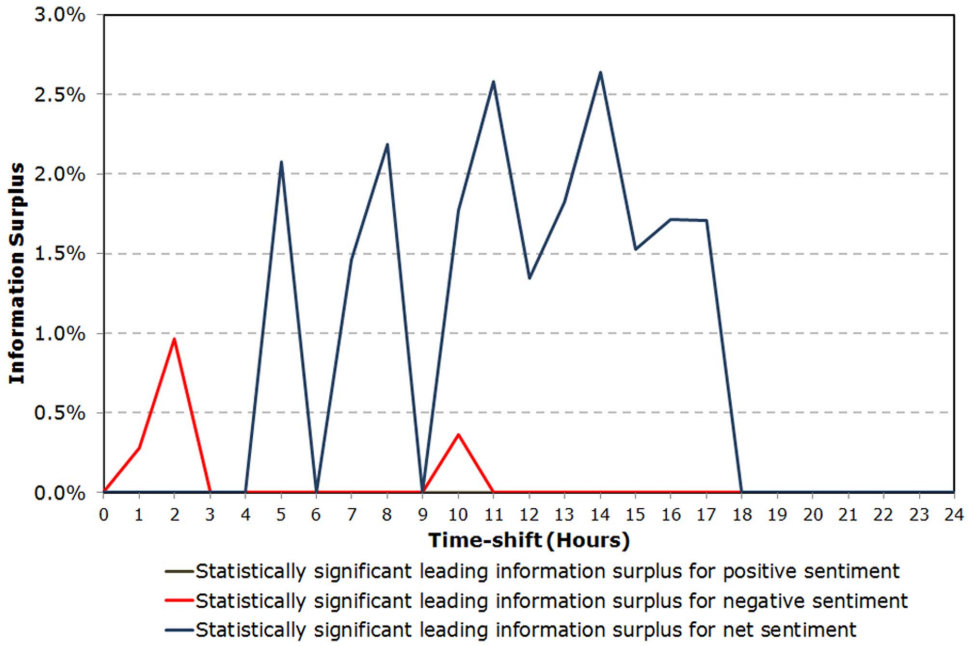

Figure 5 | *Sentiment* data can *lead* financial data for a range of *time-shifts* in a statistically-significant manner. By way of example, we demonstrate the statistically-significant *leading information surplus* between hourly changes in the *sentiment* data for the Twitter Filter: "$GOOG" AND/OR "Google" and the hourly returns of Google, Inc. CFDs. Here, we demonstrate the performances of the three different *sentiment* types (positive, negative and net), as produced by the SentiStrength classifier. Instances where the *information surplus* is positive denotes: a *leading time-shift* for which the hourly changes in the *sentiment* data contain more *information* about the security's hourly returns ahead of time than at zero *time-shift* in a statistically-significant manner and simultaneously this *sentiment* data is more *leading* than *trailing*. Thus, for such instances we can say that social media *sentiment* data does precede the financial data. Note that for the financial-instrument/Twitter-Filter combination shown in this example, there are no instances where hourly changes in the positive *sentiment*s of the Tweets performed successfully in *leading* the security's hourly returns. However, there are three instances where hourly changes in the negative *sentiment* component of the Tweets do *lead* the security's hourly returns with a confidence interval of 99%. Similarly, we observe eleven instances in this example where hourly changes in the net *sentiment* component of the Tweets *lead* the security's hourly returns in a statistically-significant manner.

determine that of the twelve financial-instrument/Twitter-Filter combinations we admit from our *sentiment analysis* experiments (as listed in Table 2), only three assets show statistically-significant instances of hourly changes in Tweet *volumes* being able to *lead* securities' hourly returns. These are: The Home Depot, Inc. CFDs with messages filtered by the Company Name AND/OR the Ticker-ID Twitter Filter (largest *information surplus* of 2.016% at a *leading time-shift* of 15-hours); Apple, Inc. CFDs with messages filtered solely by the "$AAPL" Ticker-ID Twitter Filter (largest *information surplus* of 0.981% at a *leading time-shift* of 2-hours); and J.P. Morgan, Inc. CFDs with messages filtered by the Company Name AND/OR the Ticker-ID Twitter Filter (largest *information surplus* of 1.213% at a *leading time-shift* of 13-hours). We do however identify one additional case (Bank of America, Corp. CFDs with messages filtered by Company Name AND/OR the Ticker-ID Twitter Filter) for which hourly changes in Tweet message *volumes lead* the security's hourly returns with an *information surplus* of 0.607% at a *leading time-shift* of 1-hour, but hourly changes in Tweet *sentiments* do not. When considering the ability of $\Delta_{Tweet\ volume}$ to *lead* $|\Delta_{Price}|$ (the absolute hourly returns), we determine that of the twelve financial-instrument/Twitter-Filter combinations we admit from our *sentiment analysis* experiments (as listed in Table 2), only four assets show statistically-significant instances of hourly changes in Tweet *volumes* being able to *lead* the securities' absolute hourly returns. These are: The Home Depot, Inc. CFDs with messages filtered by the Company Name AND/OR the Ticker-ID Twitter Filter (largest *information surplus* of 2.232% at a *leading time-shift* of 15-hours); Apple, Inc. CFDs with messages filtered solely by the "$AAPL" Ticker-ID Twitter Filter (largest *information surplus* of 0.944% at a *leading time-shift* of 2-hours); J.P. Morgan, Inc. CFDs with messages filtered by the Company Name AND/OR the Ticker ID Twitter Filter (largest *information surplus* of 1.374% at a *leading time-shift* of 16-hours); and Intel, Corp. CFDs with messages filtered by the Company Name AND/OR the Ticker-ID Twitter Filter (largest *information surplus* of 0.518% at a *leading time-shift* of 2-hours). As with the experiment of $\Delta_{Tweet\ volume}$ *leading* Δ_{Price} (the hourly returns), we do however identify one additional case (Bank of America, Corp. CFDs with messages filtered by the Company Name AND/OR the Ticker-ID Twitter Filter) for which hourly changes in Tweet message *volumes lead* the security's absolute hourly returns with an *information surplus* of 0.652% at a *leading time-shift* of 14-hour, but hourly changes in Tweet *sentiments* do not.

The relative performances of the largest *statistically-significant information surplus* values seen for the Tweet *sentiment* experiment, and the two Tweet *volume* experiments are seen in Figure 2, where we demonstrate that hourly changes in social media *sentiment* carry stronger abilities to *lead* securities' returns, over and above what is available with Tweet *volume* data. We do however note that Tweet *volumes lead* assets' absolute returns ($|\Delta_{Price}|$) to a greater extent than actual returns (Δ_{Price}).

Discussion

The results of our study suggest that, for the majority of financial instruments considered, hourly changes in social media *sentiment* do not contain *lead-time information* about securities' hourly returns when evaluated from a data-set of up to 10% of all messages from Twitter's network. This is primarily driven by two limiting factors. Firstly, there is insufficient Tweet volume available on the assets we've investigated to warrant the experiment. Secondly, for some financial instruments which do attract sufficient message volumes,

Table 2 | Social media *sentiment* can *lead* financial returns. For each instrument above we show its largest statistically-significant *information surplus* seen in the study, i.e. Twitter *sentiment*'s best ability to *lead* financial data ahead of time, relative to no *time-shift*. For each instrument, we also offer a summary of: the search characteristics of their Twitter Filters; their mean minutely message volume over the 3-month collection period; and their corresponding largest statistically-significant *information surplus*. We also demonstrate the *leading time-shift* (in hours) at which this occurs, and the corresponding *sentiment* type (positive, negative or net). We also report the total number of statistically-significant instances where social media *sentiment leads* financial data. Note: as discussed in the Methods, the full 24-hour autocorrelation-removal moving mean windows have been used throughout. We observe that Twitter Filter #11 (''$AAPL'') is the only filter admitted which uses just the financial instrument's industry Ticker-ID. *: We witness unexpectedly-low hourly volumes of string-unfiltered US Tweets. This is because we employed the most-accurate location-detection methodology available: only admitting those Tweets which are stamped with geographical-coordinates encompassed within the extremes of the United States' border. The majority of Tweets are not stamped with geographical-coordinates since typically only those messages which are sent from GPS-enabled devices may contain geographical-coordinates. Nonetheless, this hourly message volume was sufficient to pass our minimum mean message volume threshold of 1 message per minute. Finally, we note that our methodology identifies the following financial-instrument/Twitter-Filter combinations as inadmissible due to a lack of statistical-significance: Microsoft CFDs, FTSE100 CFDs and Futures, S&P500 CFDs, IBM CFDs, Wal-Mart CFDs and Bank of America CFDs. These assets do attract sufficient Tweet volumes, but their *sentiments* are not able to *lead* financial data in a statistically-significant manner for any of the *leading time-shifts* considered in this investigation

#	Instrument name	Twitter Filter	Mean message volume per minute	Largest statistically-significant *information surplus*
1	Apple, Inc. CFDs	$AAPL AND/OR ''Apple''	126.7	0.140%
2	Amazon.com, Inc. CFDs	$AMZN AND/OR ''Amazon''	123.1	3.473%
3	Google, Inc. CFDs	$GOOG AND/OR ''Google''	184.0	2.638%
4	Intel, Inc. CFDs	$INTL AND/OR ''Intel''	12.9	1.414%
5	Coca-Cola, Co. CFDs	$KO AND/OR ''Coca Cola'' AND/OR ''Coca-Cola''	24.8	0.723%
6	McDonald's, Corp. CFDs	$MCD AND/OR ''McDonald's'' AND/OR ''McDonalds''	46.5	1.902%
7	S&P500 Futures	String-unfiltered US Tweets	142.7*	2.462%
8	Oracle, Corp. CFDs	$ORCL AND/OR ''Oracle''	5.0	0.363%
9	Cisco Systems, Inc. CFDs	$CSCO AND/OR ''Cisco''	4.0	2.766%
10	The Home Depot, Inc. CFDs	$HD AND/OR ''Home Depot''	1.9	2.813%
11	Apple, Inc. (Ticker only) CFDs	$AAPL	1.8	3.347%
12	J.P. Morgan, Inc. CFDs	$JPM OR ''JPMorgan'' OR ''JP Morgan''	1.1	3.936%

#	Instrument name	*Leading time-shift* corresponding to the largest *information surplus*	*Sentiment* type corresponding to the largest *information surplus*	Number of statistically-significant *leading information surplus time-shifts*
1	Apple, Inc. CFDs	10	Negative	2
2	Amazon.com, Inc. CFDs	20	Net	30
3	Google, Inc. CFDs	14	Net	14
4	Intel, Inc. CFDs	1	Negative	2
5	Coca-Cola, Co. CFDs	8	Positive	13
6	McDonald's, Corp. CFDs	13	Net	7
7	S&P500 Futures	22	Net	1
8	Oracle, Corp. CFDs	1	Net	1
9	Cisco Systems, Inc. CFDs	13	Net	15
10	The Home Depot, Inc. CFDs	11	Positive	8
11	Apple, Inc. (Ticker only) CFDs	14	Negative	2
12	J.P. Morgan, Inc. CFDs	12	Positive	2

we verify that Twitter *sentiment* does not *lead* financial markets in a statistically-significant manner. These assets are: Microsoft Corp. CFDs, FTSE100 CFDs, FTSE100 Futures, S&P500 CFDs, IBM Inc. CFDs, Wal-Mart Inc. CFDs and Bank of America Corp. CFDs. In particular we note that UK string-unfiltered Tweets are not able to *lead* the hourly returns of FTSE100 Futures or CFDs, however we do identify that US string-unfiltered Tweets do demonstrate the ability to *lead* the hourly returns of S&P500 index Futures with a statistically-significant *leading information surplus* of 2.46% observed at a *leading time-shift* of 22-hours, acting in support of previous predictive[7] and correlative[8,9] social media message analysis studies.

Overall we do identify a total of twelve financial-instrument/Twitter-Filter combinations from our 10% Twitter feed dataset for which we can argue that hourly changes in social media *sentiment* do indeed contain *lead-time information* about securities' hourly returns. Ten of these represent individual stocks filtered by Company Name AND/OR Ticker-ID, one represents a stock filtered solely by its Ticker-ID (Apple, Inc. via ''$AAPL''), and one represents an index (S&P500 Futures).

To assess commonalities to these results, we first use a k-means clustering algorithm[23] configured for two categories to group the observed message volumes on companies, as seen in Table 3. We identify that Tweet volumes relating Apple Inc., Amazon.com Inc. and Google Inc. are clustered together by the k-means algorithm, and are separated from the remaining nine financial-instrument/Twitter-Filter combinations. These three firms also have the highest brand values[24] of all the firms admitted by our study. We therefore quantitatively arrive at an intuitive observation: the highest value brands attract the most Twitter message volumes.

Whilst all of the stocks admitted by our study are the most prominent global brands by value, we however do not observe a relationship between the mean per-minute Twitter message volumes for a given security or that company's brand value, and the ability for Twitter *sentiment* data to *lead* the financial data. We explain this

by the notion that a Twitter Filter mentioning a company's name (e.g. "Amazon") does not necessarily guarantee that filtered-in messages will only contain opinions on that firm. The messages may instead contain mentions of a company's service (e.g. "Check out this great deal on Amazon.com") or may in fact be entirely unrelated (e.g. "The Amazon river is unbelievably long"). We therefore wish to make a critical observation: whilst we demonstrate instances of where social media *sentiment* filtered by company name may *lead* financial markets in a statistically-significant manner, it is likely that the potential strength of such relationships is diminished by our inability to guarantee that we can filter Tweets to only allow through direct opinions on a company's future performance.

With regards to string-unfiltered Tweets from the US and the UK *leading* the hourly returns of nations' main indices, we determine that hourly changes in Twitter *sentiment* data do not appear to *lead* the hourly returns of FTSE100 index's CFDs or Futures in a statistically-significant manner. However, we do observe that the hourly changes in the *sentiments* of string-unfiltered Tweets from the US do demonstrate the ability to *lead* the hourly returns of S&P500 Futures in a statistically-significant manner (but not those of S&P500 CFDs). However we note that hourly changes in Twitter *sentiment* data only led the hourly returns of S&P500 Futures for one *time-shift* (22-hours), as seen in Table 2. Here, it is the net *sentiment* of Tweets from the US which demonstrate this ability. We argue that this is an expected result since it can be suggested intuitively that it ought to be the *overall* mood of a nation which could *lead* its main stock indices, if ever (and not solely the positive mood and/or the negative mood).

With regards to filtering Tweets solely by the industry Ticker-ID, we identify that only Apple, Inc. CFDs attract sufficient Tweet volumes to be admitted in our study. In this case, the search-term "$AAPL" resulted in a mean minutely message volume of 1.79. Such messages resulted in a peak *information surplus* of 3.34% at a statistically-significant *leading time-shift* of 14-hours, ascertained from negative *sentiments*. We also observe a similar *information surplus* of 3.28% at a statistically-significant *leading time-shift* of 15-hours. Because Tweets which can be filtered in by Apple, Inc.'s industry Ticker-ID are likely to contain direct opinions about the stock's performance, we suggest that hourly changes in the *sentiments* of such messages are intuitively more likely to *lead* the security's hourly returns than Tweets which match Apple's company name in general. We witness this in our results: the peak *information surplus* ascertained from Tweets mentioning Apple's name is only 0.14%, at a statistically-significant *leading time-shift* of 10-hours (as seen in Table 2). We also observe that in both cases, it is only the negative *sentiment* on Apple which appears to *lead* the financial data, suggesting that in the studied period, Apple, Inc.'s stock prices may respond more strongly to negative *sentiments* than positive *sentiments* or net *sentiments*.

We also demonstrate that the largest statistically-significant *information surplus* values we identify are caused by different *sentiment* types (positive, negative or net), as shown in Table 2. In such a manner we highlight that future market movements are influenced by the demographics of Twitter's users, who may Tweet predominantly positive or negative messages, depending on the company in question.

Finally, by also applying our methodology to Tweet *volumes* (rather than Tweet message *sentiments*), we demonstrate that for our dataset of up to 10% of all messages from Twitter's network, hourly changes in the *sentiments* of social media messages *lead* securities' hourly returns in a statistically-significant manner for more *time-shifts* and to a greater extent than hourly changes in Tweet message *volumes* (as shown in Figure 2). Our study admits twelve assets for which hourly changes in social media *sentiment lead* financial data. Of these, hourly changes in Twitter *volumes* occasionally led the hourly returns of three of these assets, and the absolute hourly returns of four of these assets in a statistically-significant manner. We do however identify one additional case (Bank of America, Corp. CFDs) in which hourly changes in Tweet message *volumes lead* the security's hourly returns in a statistically-significant manner whilst hourly changes in Tweet message *sentiments* do not. We can therefore conclude that *sentiments* of social media messages show consistently stronger abilities to *lead* financial markets than social media *volumes*, and we therefore argue that further attention should be given to exploring this valuable source of data.

We argue that social media *sentiment* contains *lead-time information* about financial data on S&P500 index Futures or on a narrow

Table 3 | k-means clustering of admitted assets by Tweet volume. We run a k-means[20] clustering algorithm on the mean minutely volumes of Tweets collected over the entire study for the financial-instrument/Twitter-Filter combinations for which we deem hourly changes in social media *sentiments* to *lead* securities' hourly returns in a statistically-significant manner. By clustering these volumes into two categories, we compare the mean minutely Tweet volume to the financial-instrument's brand value[24]. We observe that the companies grouped into cluster 1: Apple Inc., Amazon.com Inc. and Google Inc. (with a centroid of 144.1 messages per minute) are also the most popular brands admitted in our study. Cluster 2 encapsulates the remaining companies admitted by our study (with a centroid of 12.3 messages per minute). We therefore quantitatively show the intuitive relationship that companies of high brand-value are also represented strongly in terms of Tweet volumes, and suggest that any trading strategies built on the analytics of social media data should give particular attention to such companies due to the high-density of Tweets *: Note that we exclude message volumes attributed to the S&P500 index Futures and to Apple, Inc. CFDs (collected solely via the Ticker-ID Twitter Filter) from these clustering calculations

#	Instrument name	Twitter Filter	Mean message volume per minute	k-means clustering category for message volume	Brand value (m)
1	Apple, Inc. CFDs	$AAPL AND/OR ''Apple''	126.7	1	$87,304
2	Amazon.com, Inc. CFDs	$AMZN AND/OR ''Amazon''	123.1	1	$36,788
3	Google, Inc. CFDs	$GOOG AND/OR ''Google''	184.0	1	$52,132
4	Intel, Inc. CFDs	$INTL AND/OR ''Intel''	12.9	2	$21,139
5	Coca-Cola, Co. CFDs	$KO AND/OR ''Coca Cola'' AND/OR ''Coca-Cola''	24.8	2	$34,205
6	McDonald's, Corp. CFDs	$MCD AND/OR ''McDonald's'' AND/OR ''McDonalds''	46.5	2	$21,642
7	S&P500 Futures	String-unfiltered US Tweets	142.7	1	N/A
8	Oracle, Corp. CFDs	$ORCL AND/OR ''Oracle''	5.0	2	$16,047
9	Cisco Systems, Inc. CFDs	$CSCO AND/OR ''Cisco''	4.0	2	$15,468
10	The Home Depot, Inc. CFDs	$HD AND/OR ''Home Depot''	1.9	2	$23,423
11	Apple, Inc. (Ticker only) CFDs	$AAPL	1.8	2	N/A
12	J.P. Morgan, Inc. CFDs	$JPM OR ''JPMorgan'' OR ''JP Morgan''	1.1	2	$13,775

spectrum of highest brand-worth companies, based on our dataset of up to 10% of all messages from Twitter's network. We do not make claims to social media *sentiment* data having a causal relationship with financial data. However, we do identify instances where social media *sentiment* data contains statistically-significant indications of *leading* financial data, over and above what Twitter message *volumes* can provide. We also observe a small number of assets for which company name AND/OR Ticker-ID Twitter Filters attract a particularly large minutely message volume – such messages reference Apple Inc., Google Inc. and Amazon.com Inc., all of which are companies with the highest brand values[21]. Therefore, we suggest that any potential trading strategies based on the *sentiment* analytics of social media data should consider placing emphasis on these high message-volume companies in order to receive the highest-density "collective wisdom"[16] on a stock's potential future performance. However, we argue that messages solely matching a company's industry Ticker-ID are more likely to contain information referring just to investors' opinions on its financial performance. We observe this in the case of Apple, Inc. CFDs, whereby the *sentiments* of Tweets filtered just by "$AAPL" yield a greater statistically-significant *information surplus* about the firm's hourly returns ahead of time than messages which also match the company's name. Apple, Inc. however, is the only company considered in our study which attracts a sufficient message volume using solely an industry Ticker-ID filter, indicating that for industry Ticker-IDs to be of use in such a manner, we must hope for Twitter's popularity to rise, and therefore generate larger message volumes.

In conclusion, we suggest that when evaluating a 10% sample of all messages from a network, social media *sentiment* in a broad-based system like Twitter is indicative of future market movements only in a narrow range of assets, and that such social media *sentiments* are more indicative than just message *volumes*. We argue that, since the *sentiments* of social-media messages carry more statistically-significant *information* about future market performance than just the *volumes* of the messages themselves, such data-sources should receive further attention. We also argue that the companies for which Tweets can *lead* future market movements have a tie to the global popularities of such firms, and the demographics of those who discuss them. Whilst we do identify a number of world-famous companies on which Tweet *sentiments* appear to *lead* future financial returns in a statistically-significant manner, we argue that social media's ability to *lead* financial data could be improved if it were possible to filter in only opinions on a firm's future performance (rather than including all Tweets which mention a company's name). We therefore suggest that, if in the future financial professionals' desires to share their investment opinions through social media networks grew, then the potential for structuring successful profit-making strategies from such data sources would also increase.

Let us note that the criteria for statistical-significance that we have adopted in this study are very conservative and the dependency measure (Mutual Information, with histogram binning using Sturges' Histogram Rule) that we have adopted is not specifically fine-tuned to the purposes of the present investigation. However, we report that we have tested the robustness of our results with respect to changes in histogram binning size, finding non-significant differences. We do however suggest that is very likely that with less-restrictive and purpose-specific dependency methodologies, larger and more significant *leading* signals could be captured[25]. We also note that our approach to addressing issues of multiple-hypothesis testing with reference to the four variable types (positive, negative and net *sentiments*, as well as message *volumes*) ignores any possible overlaps in error rates. Thus, whilst we do perform our tests for statistical-significance for each variable type independent of one another, the inclusion of any corrections could multiply the p-values by a maximum of four. This would therefore lower the significance of our results, but not beyond the 95% confidence level.

We also note that our methodology considered a dataset of up to 10% of all messages from Twitter's network, and thus we did not evaluate all of the messages available through Twitter during our 3-month data collection period. We therefore suggest that the evaluation of all of Twitter's messages could have also identified larger and more significant *leading* signals. Furthermore, we do not have a basis for suggesting that Twitter message *volumes* are stationary. Therefore, we cannot argue that an estimate of their total *volume* can be compared to the estimate of the average *sentiment* of all Tweets, based on the extrapolation of the 10% Twitter feed to a theoretical 100% feed. Finally, we note that the SentiStrength *sentiment* classifier used in our study is not programmed to infer or correctly rank complex elements of human speech such as sarcasm and irony. We therefore suggest if this could be overcome, a more accurate indication of the *sentiment* of text could be ascertained.

Methods

Twitter data were collected via programmatic connection to Twitter's 10% elevated-access Gardenhose feed using a custom-coded Twitter Collection Framework (TCF). As an evolution from SocialSTORM, University College London's Social Media Collection, Processing and Analytics Engine[26], the TCF is capable of filtering Tweets based on string-filters and/or geographic-coordinate filters. The platform is integrated version 2.2 of SentiStrength[19], a highly-competent[20] and fully-transparent dictionary-based *sentiment analysis* tool developed for the accurate ranking of grammatically and lexically-incorrect English text often used in social media messages. We configured SentiStrength with the lexicon as at 16th October 2012 to its default settings without setting additional parameters. Notably, in this configuration the package takes into account the negation of text by assigning negative *sentiments* to terms which are preceded by negators such as "not".

The TCF permits real-time *sentiment analysis* of Tweets to produce the *sentiment* data used in this study, with the capacity to theoretically sustain Twitter's 100% Firehose Feed. The *sentiments* generated for this study were considered independently on three scales: 'neutral to positive'; 'negative to neutral'; and on an arbitrary emotion scale from 'very negative' to 'very positive'.

Intraday financial tick-data data (the financial data used in this study) were collected from two sources: Futures prices from Bloomberg and CFD prices from a European investment management firm accredited by its country's financial standards association.

The social media data and financial data were analysed in accordance to the processes described using a set of custom-coded MATLAB-based frameworks. Further details are given in the *Supplementary Information*.

1. Kietzmann, J. H., Hermkens, K., McCarthy, I. P. & Silvestre, B. S. Social media? Get serious! Understanding the functional building blocks of social media. *Bus. Horizons* **54**, 241–251 (2011).
2. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. Understanding the Demographics of Twitter Users. Paper presented at the *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain. Menlo Park, CA, USA: The AAAI Press. (July 2011).
3. de Vries, L., Gensler, S. & Leeflang, P. S. H. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *J. Interact. Mark.* **26**, 83–91 (2012).
4. Asur, S. & Huberman, B. A. Predicting the Future with Social Media. Paper presented at the *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Theory*, Toronto, Canada. DOI: 10.1109/WI-IAT.2010.63. (August 2010).
5. O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. Paper presented at the *Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, USA. Menlo Park, CA, USA: The AAAI Press. (May 2010).
6. Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *J. Comp. Sci.* **2**, 1–8 (2011).
7. Zhang, X., Fuehres, H. & Gloor, P. A. Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear" *Procedia Soc. Behav. Sci.* **26**, 55–62 (2011).
8. Mao, Y., Wei, W., Wang, B. & Liu, B. Correlating S&P500 stocks with Twitter data. Paper presented at the *First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, Beijing, China. New York, NY, USA: ACM. (August 2012).
9. Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. Correlating Financial Time Series with Micro Blogging Activity. Paper presented at the *Fifth ACM International Conference on Web search and Data Mining*, Seattle, USA. New York, NY, USA: ACM. (February 2012).
10. Preis, T., Moat, H. S. & Stanley, H. E. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci. Rep.* **3**, 1684 (2013).
11. Challet, D. & Bel Hadj Ayed, A. Predicting financial markets with Google Trends and not so random keywords. *arXiv preprint arXiv:1307.4643* (2013).

12. Preis, T., Reith, D. & Stanley, H. E. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philos. T. R. Soc. A.* **368**, 5707–5719 (2010).
13. Bordino, I., Battiston, S., Caldarelli, G. & Cristelli, M. Web search queries can predict stock market volumes. *PloS one* **7**, e40014 (2012).
14. Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E. & Preis, T. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Sci. Rep.* **3**, 1801 (2013).
15. Grossman, S. J. & Stiglitz, J. E. On the Impossibility of Informationally Efficient Markets. *Am. Econ. Rev.* **70**, 393–408 (1980).
16. Brody, D., Meister, B. & Parry, M. Informational inefficiency in financial markets. *Math. Fin. Econ.* **6**, 249–259 (2012).
17. Saavedra, S., Duch, J. & Uzzi, B. Tracking Traders' Understanding of the Market Using e-Communication Data. *PLoS ONE* **6**, e26705 (2011).
18. Oliveira, N., Cortez, P. & Areal, N. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. Paper presented at the *3rd International Conference on Web Intelligence, Mining and Semantics*, Madrid, Spain. New York, NY, USA: ACM. (June 2013).
19. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Imf. Sci. Technol.* **61**, 2544–2558 (2010).
20. Thelwall, M., Buckley, K. & Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Imf. Sci. Technol.* **63**, 163–173 (2012).
21. Dionisio, A., Menezes, R. & Mendes, D. A. Mutual information: a measure of dependency for nonlinear time series. *Phys. A.* **344**, 326–329 (2004).
22. Sturges, H. A. The Choice of a Class Interval. *JASA* **21**, 65–66 (1926).
23. MacQueen, J. B. Some Methods for Classification and Analysis of MultiVariate Observations. Paper presented at the *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA. Berkeley, CA, USA: University of California Press. (June 1965).
24. D'Souza, S. Brandirectory Global 500 2013 Top Brands. *Brandirectory* (2013). Available at: http://brandirectory.com/league_tables/table/global-500-2013. Accessed: 29/Oct/2013.
25. Zaremba, A. & Aste, T. Measures of Causality in Complex Datasets with application to financial data. arXiv preprint arXiv:1401.1457 (2014).
26. Wood, R., Zheludev, I. & Treleaven, P. Mining Social Data with UCL's Social Media Platform. Paper presented at the *2012 International Conference on Data Mining*, Las Vegas, NV, USA. Las Vegas, NV, USA: CSREA Press. (July 2012).

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council of the United Kingdom.

Author contributions

I.Z. developed the design of the study, performed analyses, wrote the main manuscript text, discussed the results, and prepared all figures and tables. R.S. and T.A. guided the development of the study suggesting analyses and methodologies.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zheludev, I., Smith, R. & Aste, T. When Can Social Media Lead Financial Markets? *Sci. Rep.* **4**, 4213; DOI:10.1038/srep04213 (2014).