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 messages 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, 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 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, motion picture box office returns and election outcomes. Of present focus is the prediction of financial markets via the analysis of Tweets and other comparable data sources such as Google Trends, Yahoo! search engine data and Wikipedia articles. Whilst the rationales behind all these analyses are united together by the existence of information inefficiency in financial markets, 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 messages, search engine queries, and Wikipedia views & edits. 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 sentiment of social media messages. Such methodologies, such as the work by Zhang et al. which considers up to 1% of all Tweets, attempt to anticipate markets ahead of time 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. 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 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 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”           |
| American Express, Co. CFDs    | Ticker ID AND/OR Company Name           | $AMZN                           |
| 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” |
| 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           | $INJ 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” |
| Coca-Cola, Co. CFDs           | Ticker ID AND/OR Company Name           | $KO AND/OR “Coca-Cola” AND/OR “Coca Cola” |
| McDonald’s, Corp. CFDs        | Ticker ID AND/OR Company Name           | $MCD AND/OR “McDonald’s” AND/OR “McDonalds” |
| 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 AND/OR 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”               |
| 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                            |

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. “SCSCO”; 2) combined instrument Ticker ID and/or Company Name filters, e.g. “SCSCO” 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, a leading research-oriented, 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, and more recently in ranking the sentiments of YouTube video comments, Tweets, and online posts on the Runner’s World forum. 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 were calculated relative to the previous time-window. Similarly, the hourly changes in the financial prices were calculated relative to the previous hour to generate an indication of hourly returns. We thus compare the Δsentiment for each financial-instrument/Twitter-Filter combination.

Determination of 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 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, 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 \( \mu \) 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, \( \mu_{\text{increase}} = y - x = \mu_{L>0} - \mu_{L=0} \) as the information surplus. If the information surplus is positive, i.e. \( \mu_{\text{increase}} > 0 \), 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. \( \mu_{\text{increase}} < 0 \), 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 permuted 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 time-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.
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. 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-1D-only Twitter Filters: “SAMZN” (Amazon.com Inc.), “ST” (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.), “$SINTC” (Intel Corp.), “$JPM” (J.P. Morgan Inc.), “$SMCD” (McDonald’s Corp.), “$MSFT” (Microsoft Corp.) and “$ORCL” (Oracle Corp.).

Finally, we also reject the following Twitter Filters which use Company Names AND/OR industry ticker-1Ds, 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 autocorrelation-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 permute 10,000 times the hourly changes in sentiment data for each sentiment type: positive, negative, or net against the hourly changes in asset price data 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 $\Delta_{price}$ (the hourly
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 and search engine message volumes with financial market performance. We then also repeat this experiment to consider $\Delta_T$ against $|\Delta_{F_C}|$ (the absolute hourly returns) to further explore the ability of hourly changes in Tweet volumes to lead securities' hourly returns. We

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 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.

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.
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 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 the 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 \( \Delta_{\text{Tweet \ volume}} \) to lead \( \Delta_{\text{Price}} \) (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); and Intel, Corp. 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 \( \Delta_{\text{Tweet \ volume}} \) leading \( \Delta_{\text{Price}} \) (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 \( (\Delta_{\text{Price}}) \) to a greater extent than actual returns \( \Delta_{\text{Price}} \).

**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,
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 CFDs. In particular we note that US 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 and correlative 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\(^*\) 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\(^*\) 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

| # | 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 “J P Morgan” | 1.1 | 2 | $13,775 |
and more significant restrictive and purpose-specific dependency methodologies, larger respect to changes in histogram binning size, finding non-significant we report that we have tested the robustness of our results with Sturges’ Histogram Rule) that we have adopted is not specifically adopted in this study are very conservative and the dependency ing strategies from such data sources would also increase. We therefore suggest that, if in the future financial professionals’ (rather than including all Tweets which mention a company’s name). possible to filter in only opinions on a firm’s future performance information sentiments Tweets can of the messages themselves, such data-sources should volumes. 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 indepen- denty 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 col- lected 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.

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 Annotation Engine[22] the TCF is capable of filtering Tweets based on string-filters and/or geographic-coordinate filters. The platform is inte-grated version 2.2 of SentiStrength[19], a highly-competent 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 indepen-dently 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. J. 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., Balasubramanayan, 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., Fuhres, 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., Giornis, A. & Jaime, 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. & Uzi, 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. Inf. Sci. Technol.* **61**, 2544–2558 (2010).
20. Thelwall, M., Buckley, K. & Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.* **63**, 163–173 (2012).
21. 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
IZ. 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).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0