Paper Title: Bull Bear Balance: A Cluster Analysis of Socially-informed Financial Volatility

Name: Derek Lukacsko

Paper Type: conference proceeding

Contributions: I wrote the abstract and the majority of the paper aside from the introduction. The methodology is shared between co-author Jonathan Manfield and me. I helped develop the open-source Python repository now available on Github.

Bull Bear Balance: A Cluster Analysis of Socially Informed Financial Volatility

Jonathan Manfield
Computer Science Department
University College London (UCL)
j.manfield@cs.ucl.ac.uk

Derek Lukacsko
Computer Science Department
University College London (UCL)
derek.lukacsko.15@ucl.ac.uk

Tharsis T. P. Souza
Computer Science Department
University College London (UCL)
t.souza@cs.ucl.ac.uk

***Abstract*—The use of alternative data in financial applications has gained momentum in recent years with the increased availability of data along with computational resources. While traditional financial pricing theory supports an efficient market hypothesis, recent research has shown that data mining of exogenous feeds can provide further information to inform market activity. Social media has become an increasingly important source of this information due to its abundant, directed, and real-time nature. However, little is known about what combination of social media and financial features is indicative of market activity. In this work, we investigate what combination of social media and financial features are present when social media data is effective for reducing uncertainty about future stock volatility. Moreover, identification of feature profiles from clusters of stocks indicates that sentiment polarity (i.e. positive or negative) taken alone is not enough to infer future volatility, instead a balance of bullish and bearish signals are preferred even above commonly identified features in the literature such as message volume and market cap. This is important because by combining bullish and bearish sentiment and a range of other social and financial variables we are able to generate a time series which is more informative about volatility than any of the individual feature time series. Robustness of these findings is verified across 500 stocks from both NYSE and NASDAQ exchanges. Reported results are reproducible via an open source library for social-financial analysis made freely available.**

***Keywords*—data mining; sentiment analysis; stock market; k-means clustering; mutual information; information theory; volatility**

I. Introduction

Recent research has dampened the consensus about the unpredictability of market prices; literature has shown that there are indicators in the market that can account for a portion of the noise in the market that can now be uncovered using machine learning techniques [1]. As a result, predicting the stock market has garnered the attention of diverse fields and domains where different information channels have been explored, such as news [2], [3], [4], search engines [5], [6], [7], [8] and, more recently, social media [9], [10], [11], [12].

Social media is of particular interest due to the high volume and velocity of activity of an ever-evolving network that is constantly providing and creating information. It gives real-time information that can be attributed to specific people, events, markets, and securities. This is channelled through the use of so-called *cashtags* (e.g., '$AAPL') in messages, enabling the creation of feeds associated with particular stock symbols. The use of cashtags has since propagated onto Twitter, thus providing a means to direct social sentiment to the specific stock or securities of which it references. As a result, social media user sentiment can be harnessed as a source of information in respect to financial market activity.

Now, the opinion of traders, professional bloggers, and analysts along with lay-person's opinions are aggregated in a dynamic social network that can possibly explain some of the variance in market behavior. Initial research [13] sought to quantify a relationship between social media analytics with financial market data such as daily returns. Excitingly, observed results outperformed baseline trading strategies, providing evidence that Tweets' volume can reduce uncertainty about financial returns. However, volume-based methods disregard any possible predictive information from *qualitative* aspects of the data (i.e., the actual content or content polarity). By applying sentiment analysis techniques to a corpora of tweets, a sentiment score or emotion classification can be derived to quantify this qualitative dimension [14]. The effectiveness of this semantic approach has been examined in [9], where collective moods analysed from large collections of daily tweets were used to increase an existing financial predictor performance to an accuracy of 87.6%.

Whilst a broad analysis across several financial securities might unveil that social signs are relevant to explain financial dynamics to some extent [9], [10], little is known about confounding factors that distinguish assets with predictive social signs from assets with no extra information provided by social media. Leveraged by a non-parametric analysis founded in information-theoretic measures we show that social signs can be useful for most stocks from both NYSE and NASDAQ exchanges. This is alone an interesting and very sound result compared to current literature but we extend this analysis to provide, for the first time, possible explanations of features that might be essential to distinguish predictive social signs from non-predictive ones.

A. Research Questions

- **RQ1.** ***Which stocks from NASDAQ and NYSE exhibit a significant information surplus when using social media as a leading indicator of the stocks' future volatility?***
  For each stock, information between financial and social media data is quantified in an *ex-ante* configuration.
- **RQ2.** ***Under what configuration of social media***

and financial variables are social media analytics informative of future financial movements?
We aim to determine the feature profile (using financial and social media variables) of companies that are most indicative of a statistically-significant lead-time social media information.

B. Contributions

- Identification of social media and financial features that coincide with a cluster of stocks that contain high information surplus (i.e. a lagged time series of social media data is predictive of subsequent stock volatility). The noteworthy feature in this cluster is the ratio of bullish to bearish messages, which implies that the combination of polarity in message sentiment is more important than the amount of messages generated in using social media to reduce uncertainty about a stock's volatility.
- A free, open-source package[1] where the methodology is computationally formulated into functions for the purpose of replication and further work.

II. Methodology

A. Pre-processing

1) Data: Reported results were obtained from freely available data. Social media data are provided by PsychSignal, who operates a customised Twitter and StockTwits collection framework tracking messages containing *cashtags* (e.g., $AAPL). State-of-the-art natural language processing algorithms are applied to relevant messages, labelling each with a sentiment disposition (i.e., bullish or bearish) and a measure of disposition intensity. A daily aggregate of this data is provided for each tracked stock.

For each stock with available social media data, we also consider historic records of daily financials.

- **DS1. PsychSignal Social Media Database** 7,120,506 records containing daily aggregates for 11,444 stocks[2]
- **DS2. Google Finance** Daily Market Quotes[3]

Table I. Data Configuration

Dimension	Value
Start date	01-01-2012
End date	01-01-2016
Exchanges	{NASDAQ, NYSE}
Stocks	Top 250 by largest market capitalization.

2) Volatility: We use a value of Daily True Range (see Equation 1) as a measure of financial volatility. A log transformation of TR_t is utilised to account for oscillations between different valued stocks. A time series of volatility data for each stock is derived from this calculation using financial quotes.

$$TR_t = \max[(\text{High}_t - \text{Low}_t), (\text{Low}_t - \text{Close}_{t-1}), (\text{High}_t - \text{Close}_{t-1})] \quad (1)$$

3) Dimensionality Reduction: Principal component analysis (PCA) is a dimensionality reduction technique that we utilize for feature extraction. PCA permits the reduction of numerous correlated, co-linear variables to a component (or feature set of components). By applying PCA to a set of social media variables, we obtain a time series that contains the majority of underlying information from the original features. This series is used as the input of social media data in calculating mutual information between social media and volatility.

4) Variables Analysed: Tables II and III contain the features utilised in the analyses.

Table II. Social Media Features

	Feature	Description
1	BULLISH_INTENSITY	positive sentiment polarity
2	BEARISH_INTENSITY	negative sentiment polarity
3	BULL_MINUS_BEAR	the ratio of 1 to 2
4	BULL_SCORED_ MESSAGES	positive sentiment volume, number of messages
5	BEAR_SCORED_ MESSAGES	negative sentiment volume, number of messages
6	BULL_BEAR_MSG_ RATIO	volume of bullish messages over volume of bearish messages
7	TOTAL_SCANNED_ MESSAGES	total messages, including neutral sentiment
8	LOG_BULL_RETURN	log difference in daily volume of bullish messages
9	LOG_BEAR_RETURN	log difference in daily volume of bearish messages
10	LOG_BULLISHNESS	log difference between 4 and 5
11	LOG_BULL_BEAR_ RATIO	log ratio between 4 and 5
12	LOG_BULL_MINUS_ BEAR_CHANGE	log daily difference in 3
13	TOTAL_SCANNED_ MESSAGES_DIFF	daily difference in message volume
14	TOTAL_SENTIMENT_ MESSAGES_DIFF	daily difference in volume for messages non-neutral polarity
15	PCA_SOCIAL_ CHANGE	First principal component derived from 8, 9, 10, 11, 12, 14, and 15

B. Information Surplus

The following method originates from [10] and we used it to identify stocks where social media can be used as a leading indicator of their referenced stock's volatility.

Information surplus is derived from mutual information: a measure of the mutual dependence between two feature time-series. Let S be PCA_SOCIAL_CHANGE, the daily change in the time series obtained by applying PCA and extracting the first principal component from the set of social media variables and let F be LOG_TR_DIFF, the log of the daily difference of the Daily True Range (see Equation 1). Intuitively, if the addition of series S provides information about the movements of series F, it is said that there exists a dependency or mutual information (MI) between S and F. However such a dependency is non-directional; in order to determine if S *leads* F, S must provide more information on a lagged series

[1] sentisignal package https://github.com/jonathanmanfield/sentisignal
[2] Quandl: Access PsychSignal API https://www.quandl.com/vendors/ps, Accessed: 30-09-2016
[3] Google Finance: Market Quotes https://www.google.co.uk/finance, Accessed: 30-09-2016

Table III. FINANCIAL FEATURES

	Feature	Description
16	OPEN	daily opening price
17	HIGH	daily high price
18	LOW	daily low price
19	CLOSE	daily closing price
20	VOLUME	financial volume, number of daily trades
21	LOG_RETURN	percent change in log close price
22	LOG_CLOSE	log closing price
23	LOG_HIGH	log daily high price
24	LOG_LOW	log daily low price
25	VOLATILITY_1	the absolute value of the difference between 22 and 24
26	VOLATILITY_2	the absolute value of the difference between 22 and the previous day's 21
27	VOLATILITY_3	the absolute value of the daily difference between 24 and the previous day's 21
28	TR	the max between 25, 26, and 27
29	LOG_VOLUME_DIFF	log daily difference in 20
30	LOG_TR_DIFF	log daily difference in 28

of F than the baseline MI (i.e. non-lagged). Determining if a baseline dependency exists between the series of a social media feature $S_{l=0}$ and a financial feature $F_{l=0}$ on the same day is the first step in identifying if S leads or is predictive of F. MI tells us how much the information about S reduces uncertainty of F. Equation (2) shows the formal form, where we attempt to reduce uncertainty or increase information by taking the double integral over the log of the joint entropy for both series over each distribution.

$$MI(S;F) = \int\int f(s,\mathrm{f}) \log \frac{f(s,\mathrm{f})}{f_s(s) f_{\mathrm{f}}(\mathrm{f})}\, ds\, df \qquad (2)$$

The data need to be grouped into bins to determine the mutual information between the two series. This is obligatory to calculate entropy because the probability of observing an instance i in each bin s and f constitutes the probability distributions. The number of bins is dependent on the size of the data, thus our bin sizes were typically identical per feature (i.e., there are roughly 365 daily instances of Tweets and financial data per security per year). The bin size k was calculated using Sturge's Rule (see Equation 3) which has been found to be more accurate than comparable methods when used to calculate entropy in the MI algorithm [15].

$$k = \log_2 n + 1 \qquad (3)$$

Mutual information is then computed at consecutive daily time lags. The information *gain* at time lag $l = i$ is calculated by finding the difference between mutual information at $l = i$ and $l = 0$ where i is a certain day lag and $l = 0$ corresponds to the baseline case. Information surplus is expressed as a percent of MI above what we would expect over the given time frame. Thus, if we achieve a surplus above the average MI from $l = i$ to $l = 0$, the social media time series S is said to lead F.

$$\text{Information Surplus}_l = \frac{MI(S;F)_{l=i} - MI(S;F)_{l=0}}{MI(S;F)_{l=0}} \times 100 \qquad (4)$$

C. Validating Significant Information Surplus

Surplus results are then statistically validated. This is a two-step process used to verify that information is more leading than trailing and the results are better than a randomly permuted MI.

1) We first filter companies whose surplus is more *trailing* than *leading* by identifying stocks for which the daily changes in S carry more information about the daily F in hindsight ($l = -i$) rather than during the same ($l = 0$) or a previous day ($l = i$). Intuitively, we calculate MI on a forward shift (*ex-post*) and on a backward shift (*ex-ante*), eliminating stocks where the *ex-post* MI is greater than the *ex-ante* MI. Where MI for a lag is less than MI for a retrospective advance of the series (i.e. $MI_{l=i} < MI_{l=-i}$), we can assert that $MI(S;F)$ is more trailing than leading, and thus insignificant.
2) Random permutation of the remaining symbols is then performed 100 times. With $\alpha = 0.05$, the stocks must outperform 95% of randomly permuted data to be considered to have a significant surplus.

D. Clustering

To determine the conditions under which social signs may be predictive of financial dynamics, we utilize a clustering method to determine the configuration of social media and financial variables that are indicative of a high information surplus. Each stock is represented as a vector of social media and financial variables. In addition, respective scores of information surplus and features that describe its nature (e.g., the size of the lag it was obtained at) are included. Stocks with similar configurations of social media and financial variables will be grouped together in clusters.

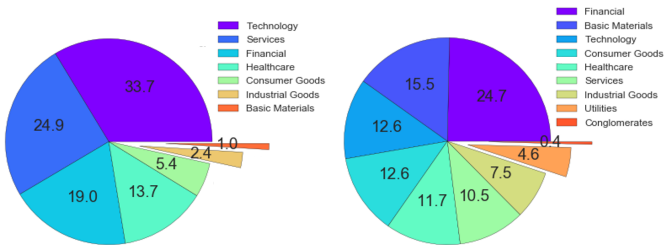
Inspecting the feature profiles of the clusters (i.e., average representation of features of constituent stocks) allows us to identify the social signs of financial dynamics that are indicative of a high surplus. The clusters that have a profile containing many significant lags and high overall information surplus will provide us insight into what configuration of variables characterizes a stock with predictable volatility.

We use k-means clustering, an unsupervised learning algorithm that partitions instances into *k* clusters by minimizing the within-cluster sum of square error (WCSS) between instances in each set S using a distance metric (see Equation 5).

$$\operatorname{argmin} S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - u_i\|^2 \qquad (5)$$

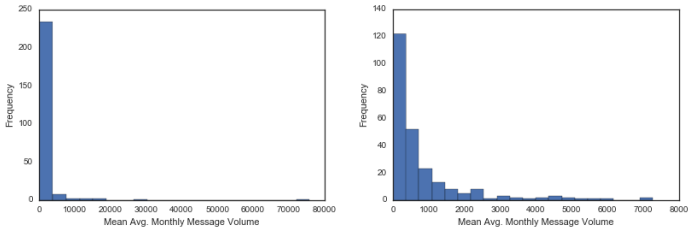
III. DESCRIPTIVE STATISTICS

The following descriptive statistics provide context to our results. Figure 1 presents the sector breakdown of top 250 stocks by market cap for both the NASDAQ and NYSE (i.e. 500 stocks in total). In context of this paper, it is interesting to note that the NASDAQ is typically characterized as being more volatile than the NYSE [16].



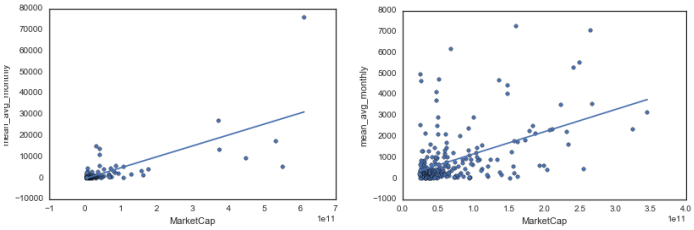
(a) NASDAQ (b) NYSE

Figure 1. **Top 250 Sector Breakdown.** Percentage breakdown by sector highlights both that the majority of large cap stocks belong to technology and financial sectors and that the NYSE has a more balanced distribution.



(a) NASDAQ (b) NYSE

Figure 2. **Probability Distribution of Tweet Volume.** There is a wide discrepancy between NYSE and NASDAQ Tweet volume for the top 250 stocks. The NASDAQ contains several outliers (most notably, $APPL) that skew the distribution to the high end. The NASDAQ contains a larger amount and a more variable distribution of Tweets with a total of 324,239 and a standard deviation of 5,489 in contrast to 212,368 and 1,241 for the NYSE.



(a) NASDAQ (b) NYSE

Figure 3. **Market cap and Tweet Volume Positive Relationship .** Top 250 stocks for NYSE and NASDAQ exhibit moderate positive correlation between market capitalization (x) and average monthly Tweet volume (y).

Fig. 2 and 3 compare the relationship between the *volume* of Tweets regarding exchange-specific securities and the size of those securities. In summary, there is a strong right skew in Tweet volume which contributes to a moderately positive relationship between the size of a security and its interest to investors as quantified by Tweet volume. In both exchanges, there exist several notable outliers such as $AAPL that contains a disproportionate volume of Tweets.

In Fig. 4 and 5, we present a range of plots related to the feature behavior of the data utilised for calculating MI. The correlation plot between all features (see tables II and III) in Fig. 4 reinforces the use of PCA dimensionality reduction. As the features are correlated, it is useful to inject a compression of these variables into our Mutual Information calculation. The correlation of social media features with other social media features is found pertaining to sentiment polarity, volume of

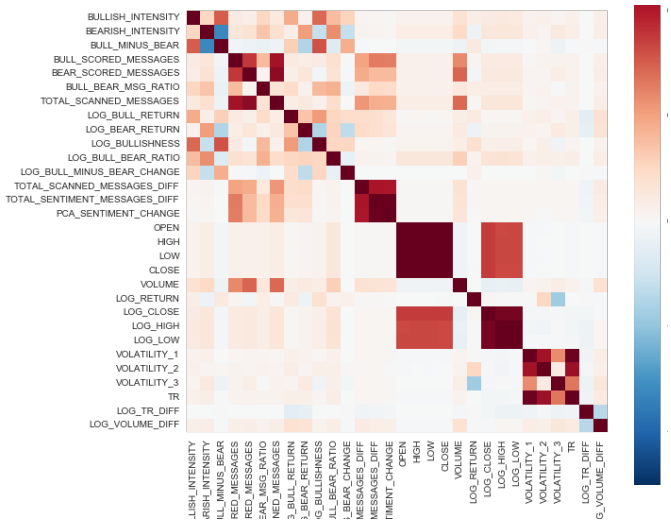
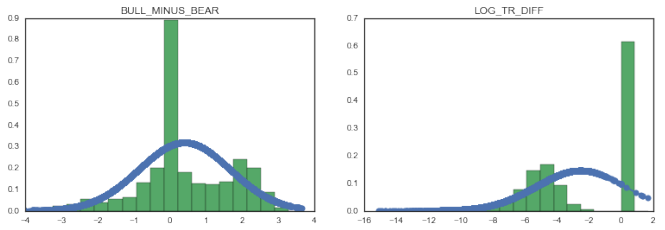


Figure 4. **NASDAQ correlation Matrix.** Upper right (and low left) quadrant reveals weak to no correlation between social and financial features. The NYSE features exhibit comparable correlation.



(a) (b)

Figure 5. **Probability Distribution Examples with Fitted Normal Curve.** The plots are exemplary of the non-normal distributions found in both our social media (a) and financial features (b) for the NASDAQ and NYSE.

messages and daily financial features, for instance.

The majority of features is also characterized by non-normal distributions (see Figure 5). For those features highly dependent on market capitalization (e.g., log returns or mean average monthly message volume as (see Figure 2)), there exists log-normal behavior. Furthermore, NASDAQ and NYSE stocks exchanges exhibit comparable results.

IV. Results

We used the earlier outlined information surplus method [10] to determine which stocks in the NYSE and NASDAQ that on average exhibit a significant *leading* information surplus using data from 01-01-2012 to 01-01-2016. We build on this method by clustering stocks and examining which configuration of variables are indicative of a high information surplus.

A. RQ1: Reducing Uncertainty about Volatility

Experiments on the NASDAQ and NYSE were carried out in tandem and produced comparable results. We found 101 stocks from the NASDAQ that exhibit a leading information surplus when using social media as an indicator of the daily change in True Range (our measure of volatility, see Equation

Table IV. EXPERIMENTAL CONFIGURATION

Application	Feature set
Mutual Information	{log true range (**29** in Table III), PCA sentiment change (**15** in Table II)}
K-means	{max. surplus perc., max. MI, position of optimal lag (1-10), market cap., log returns, bull minus bear sentiment intensity, total messages, volume of trades, log bullish intensity, number of positive lags, log true range}
Radar plot	{max. surplus perc., log true range, volume of trades, log returns, total messages, bull minus bear sentiment intensity, market cap., number of positive lags, position of optimal lag (1-10), max. MI}

1). Of the original 250 stocks examined, 149 did not have a *significant* surplus meaning the surplus in each time lag over a ten day period did not exceed the *expected* surplus. Thus for example, the periods in Fig. 6 where the blue *ex-ante* series is below the *average ex-post* for the ten day period *do not* contain a *significant* leading surplus. All 101 stocks with a significant leading surplus however passed through the second validation test by performing better than 95% of randomly permuted data.

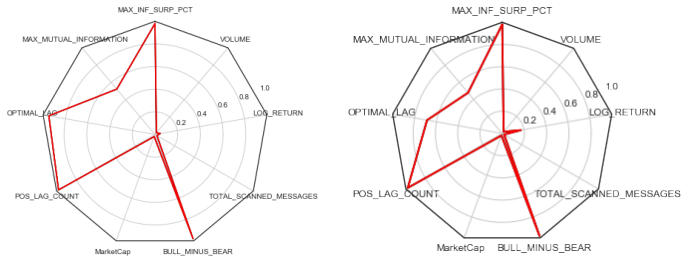
For the NYSE, 91 of the original 250 stocks exhibit a significant leading surplus. Only one company did not pass the second validation test.

A link to the full list of significant stocks from the NASDAQ and NYSE is in Appendix A.

Figure 6. **Sample Information Surplus.** For each lag, the surplus is calculated as the amount of MI above the base case, $MI_{t=0}$. A *significant ex-ante* surplus occurs where the surplus is greater than the average *ex-post* surplus for the 10-day window.

B. RQ2: Leading Surplus Indicators

Stocks with a statistically significant information surplus were clustered using relevant numeric features (see Table IV). The aim of the clustering is to identify the cluster *feature profile* of stocks with a surplus exhibited at many time lags (i.e. high `POS_LAG_COUNT` values) and the stocks with the highest surplus (i.e. high `MAX_INF_SURP_PCT` values). Indeed, the purpose of the unsupervised method is to identify the cluster profile that contains high feature values of the two aforementioned features. Results were robust across different values of number of clusters k tested. We found moderate choices of k *all* produced the noteworthy cluster profile shown in Fig. 7 where a significant surplus is almost completely contained. The two cluster profiles presented in Fig. 7 correspond to the cluster centroid feature profiles of both the NASDAQ and NYSE which contained the highest information surplus values.



(a) NASDAQ (b) NYSE

Figure 7. **Radar Plot of Cluster Centroid Feature Profiles.** We observe in both the NASDAQ and NYSE a cluster that contains the majority of leading surplus lags (`POS_LAG_COUNT`) and maximum surplus (`MAX_INF_SURP_PCT`) for $k = \{2, 3, 4, 5, 6, 7\}$. The important relationship represented in this cluster is the high value of `BULL_MINUS_BEAR`.

Interestingly, `BULL_MINUS_BEAR` is maximised in the cluster, indicating that stocks with many leading surpluses during the ten day period (`POS_LAG_COUNT`) and large significant surpluses (`MAX_INF_SURP_PERC`) also have a particular ratio of bullish and bearing intensity. This suggests that in isolation, messages with a strong positive or negative sentiment are not as informative of future volatility; rather, the *combination* of features is important in context of volatility or daily risk. Although we can speculate about the underlying factors contributing to this behavior, we note that this feature relationship seems to contradict the traditionally held assumption that *negative* sentiment is more indicative of volatility. It may be expected that a volatile security is lead by historically *negative* or *bearish* Tweets because volatility is associated with risk and high risk is synonymous with negative sentiment, but our clustering suggests otherwise; the polarity between all sentiment-loaded Tweets is important in predicting volatility. Another noteworthy implication of the clustering is the absence of `VOLUME` (number of trades), `LOG_RETURN`, and `TOTAL_SCANNED_MESSAGES` (volume of messages) in the cluster profile associated with significant leading surpluses. Those are features commonly-used in the literature that instead here did not show to be as relevant as the ratio of bullish and bearish messages.

V. CONCLUSION

Our results show that signals from social media can lead daily financial volatility in a large proportion of the 500 stocks examined from the NASDAQ and NYSE. A total of 101 (40%) stocks from the NASDAQ and a further 91 stocks (36%) from the NYSE exhibited a statistically significant information surplus. This was found by identifying an increase in mutual information (*MI*) between social and financial time series.

Whilst our framework for this section of the experiment is closely aligned with earlier works [10], our contributions are novel. Firstly, by identifying an information surplus in a large number of stocks, we have found that social media has the capability to lead financial markets in a much larger segment of the market than in previous works, which only reported 12 stocks [10]. In addition, we have broken down barriers which suggest that this type of predictive capability is reserved only

to Technology stocks [17], by reporting significant results for stocks from all included sectors.

A key aim of this paper was to go beyond the determination of the predictive capability of social media and attempt to figure out with what configuration of features does this occur. Our results showed that stocks with the highest net sentiment polarity also had the highest information surpluses. Interestingly, in contrast to other works, stocks with a high information surplus did not require a high volume of messages and did not have a high average log-return. To identify this, we characterised each stock using a average representation of social and financial variables and applied a clustering algorithm. We then inspected the group of stocks that exhibited the highest maximum information surplus.

To summarise, we have challenged the notion of the efficient market hypothesis by examining the effect of the continuously evolving source of information embedded in social media. Using a method rooted in information theory, we have presented results that have identified a large set of stocks for which social media can be informative of volatility. By clustering stocks based on joint feature sets of social and financial variables, our research has taken an important first step in characterising the conditions in which this can be the case. Results indicate that social media is most informative about financial market volatility when the ratio of bullish and bearish sentiment is high even when the number of messages is low.

ACKNOWLEDGMENT

This work was supported by PsychSignal who provided social media sentiment analytics data. We also thank Prof. Tomaso Aste for valuable comments provided.

APPENDIX A
OPEN SOURCE PACKAGE

The functionality of the open source package[4] enables social-financial analytics development in Python. Tools are included for the following purposes:

- web scraping historic financial records
- data fusion of social-financial databases
- generation of statistical time series
- measuring information surplus
- statistical significance testing
- cluster analysis
- data visualisation
- two lists of signifcant stocks from NASDAQ and NYSE

[4]sentisignal package https://github.com/jonathanmanfield/sentisignal

APPENDIX B
SIGNIFICANT STOCKS TABLE (NASDAQ)

Table V contains the NASDAQ stocks that exhibit a significant leading surplus. The symbols are organized alphabetically and the columns include the maximum surplus expressed as percent of MI above the baseline exhibited over the ten-day window. The max lag is the day when this maximum occurs (e.g. -7 = one week prior). The average surplus is the expected surplus over the ten day window. Each day with a significant leading surplus greater than 0 is tallied in the count column. The final column is the sector of the corresponding stock.

REFERENCES

[1] Bo Qian and Khaled Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, 2007.

[2] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[3] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[4] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Sci. Rep.*, 3, 2013.

[5] Chester Curme, Tobias Preis, H. Eugene Stanley, and Helen Susannah Moat. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 2014.

[6] Tobias Preis, Daniel Reith, and H. Eugene Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, 2010.

[7] Tobias Preis, Helen S. Moat, and H. Eugene Stanley. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, 3, April 2013.

[8] Zhi Da, Joseph Engelberg, and Pengjie Gao. In search of attention. *The Journal of Finance*, 66(5):1461–1499, 2011.

[9] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[10] Ilya Zheludev, Robert Smith, and Tomaso Aste. When can social media lead financial markets? *Sci. Rep. Scientific Reports*, 4, 2014.

[11] Thársis T. P. Souza, Olga Kolchyna, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis applied to finance: A case study in the retail industry. In Gautam Mitra and Xiang Yu, editors, *Handbook of Sentiment Analysis in Finance*, chapter 23. 2016.

[12] Thársis T. P. Souza and T. Aste. A nonlinear impact: evidences of causal effects of social media on market prices. arXiv preprint. http://arxiv.org/abs/1601.04535, 2016.

[13] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM, 2012.

[14] Olga Kolchyna, Thársis T. P. Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. In Gautam Mitra and Xiang Yu, editors, *Handbook of Sentiment Analysis in Finance*, chapter 5. 2016.

[15] Philip A Legg, Paul L Rosin, David Marshall, and James E Morgan. Improving accuracy and efficiency of registration by mutual information using sturges histogram rule. *Proc. Med. Image Understand. Anal*, pages 26–30, 2007.

[16] Edward A Dyl and George J Jiang. Valuing illiquid common stock. *Financial Analysts Journal*, 64(4):40–47, 2008.

[17] G William Schwert. Stock volatility in the new millennium: how wacky is nasdaq? *Journal of Monetary Economics*, 49(1):3–26, 2002.

Table V: NASDAQ Significant Leading Companies

	Symbol	Max_Surplus	Max_Lag	Avg_Surplus	Count	Sector
1	AAL	5.98	-7	0.60	1	Transportation
2	ACGL	16.46	-10	5.86	6	Finance
3	ACWI	76.02	-10	8.78	2	n/a
4	ADBE	23.44	-1	2.34	1	Technology
5	AFSI	0.36	-1	0.04	1	Finance
6	AKAM	15.02	-1	1.50	1	Miscellaneous
7	ANSS	2.54	-1	0.25	1	Technology
8	ASML	20.80	-1	2.08	1	Technology
9	BBBY	30.15	-1	3.01	1	Consumer Services
10	BUFF	8.83	-3	0.88	1	Consumer Non-Durables
11	CA	149.77	-7	44.00	7	Technology
12	CASY	9.28	-1	0.93	1	Consumer Durables
13	CDNS	2.66	-10	0.34	2	Technology
14	CG	17.93	-2	4.07	3	Finance
15	CINF	22.04	-7	3.01	2	Finance
16	CMCSA	6.33	-1	0.63	1	Consumer Services
17	CME	16.55	-1	1.65	1	Finance
18	COST	16.93	-4	2.84	3	Consumer Services
19	CSAL	21.08	-10	2.14	2	Consumer Services
20	CSGP	22.04	-1	3.57	4	Miscellaneous
21	CTAS	28.09	-1	3.81	2	Consumer Non-Durables
22	CTXS	2.61	-1	0.26	1	Technology
23	DISCB	9.46	-3	1.17	2	Consumer Services
24	DISCK	0.72	-2	0.07	1	Consumer Services
25	DOX	52.25	-7	9.85	5	Technology
26	EA	16.98	-1	1.70	1	Technology
27	ERIE	3.85	-9	0.38	1	Finance
28	EWBC	34.58	-3	12.88	6	Finance
29	FANG	3.40	-8	0.59	2	Energy
30	FB	18.25	-1	1.82	1	Technology
31	FFIV	22.07	-1	2.21	1	Technology
32	FISV	4.59	-10	0.46	1	Technology
33	FITB	26.25	-6	5.68	3	Finance
34	FOXA	13.71	-1	1.37	1	Consumer Services
35	FTNT	20.34	-1	2.03	1	Technology
36	GLPI	5.94	-8	1.51	3	Consumer Services
37	GNTX	1.81	-1	0.18	1	Capital Goods
38	HAS	30.73	-7	14.56	6	Consumer Non-Durables
39	HDS	11.63	-1	1.33	2	Consumer Services
40	HOLX	1.81	-1	0.18	1	Health Care
41	HSIC	20.18	-1	3.47	2	Health Care
42	IBKR	2.55	-1	0.26	1	Finance
43	INFO	108.15	-7	32.00	4	Technology
44	INTU	4.26	-5	1.01	4	Technology
45	JBHT	12.73	-1	1.27	1	Transportation
46	JD	38.70	-5	12.18	6	Consumer Services
47	JKHY	20.47	-8	4.89	6	Technology
48	KLAC	5.28	-1	0.53	1	Capital Goods
49	LBRDA	14.82	-1	1.57	2	Consumer Services
50	LBTYA	8.17	-1	0.82	1	Consumer Services
51	LBTYB	40.41	-7	12.91	7	Consumer Services
52	LBTYK	19.43	-7	3.04	2	Consumer Services
53	LILA	35.42	-10	3.54	1	Consumer Services
54	LILAK	39.67	-7	8.16	3	Consumer Services
55	LVNTA	26.84	-1	4.91	4	Consumer Services
56	MAR	27.35	-7	5.02	2	Consumer Services

57	MIDD	11.30	-1	1.50	2	Technology
58	MNST	13.12	-2	1.31	1	Consumer Non-Durables
59	MSCC	25.80	-7	5.98	5	Technology
60	NTAP	26.26	-1	2.63	1	Technology
61	NTRS	11.51	-1	1.15	1	Finance
62	NWS	44.64	-4	11.26	5	Consumer Services
63	NWSA	3.65	-1	0.36	1	Consumer Services
64	ON	24.65	-2	4.50	2	Technology
65	ORLY	21.53	-1	2.15	1	Consumer Services
66	PACW	4.40	-8	0.44	1	Finance
67	PAYX	21.76	-1	2.18	1	Consumer Services
68	PBCT	17.27	-4	3.10	3	Finance
69	PDCO	34.06	-1	3.93	2	Health Care
70	PPC	48.37	-5	18.08	5	Consumer Non-Durables
71	PYPL	0.15	-9	0.01	1	Miscellaneous
72	QCOM	21.89	-1	2.19	1	Technology
73	QGEN	38.94	-1	7.80	3	Health Care
74	QVCA	17.89	-4	6.90	5	Consumer Services
75	RYAAY	4.68	-6	0.55	2	Transportation
76	SABR	35.74	-1	3.57	1	Technology
77	SBAC	17.38	-1	3.00	2	Consumer Services
78	SCZ	42.50	-7	8.77	5	n/a
79	SIVB	30.48	-6	7.09	4	Finance
80	SNH	41.11	-5	4.50	2	Consumer Services
81	SNPS	13.11	-1	1.31	1	Technology
82	SSNC	18.27	-1	2.93	3	Technology
83	STLD	11.49	-6	1.15	1	Basic Industries
84	SYMC	1.92	-1	0.19	1	Technology
85	TEAM	219.09	-10	76.32	4	Technology
86	TFSL	2.09	-1	0.21	1	Finance
87	TROW	2.61	-8	0.52	3	Finance
88	TSCO	0.26	-1	0.03	1	Consumer Services
89	UHAL	17.33	-9	1.73	1	Consumer Services
90	ULTI	24.29	-1	3.35	3	Technology
91	VCIT	69.82	-8	13.20	4	n/a
92	VIP	9.70	-4	0.97	1	Public Utilities
93	VRSK	7.91	-6	0.79	1	Technology
94	VRSN	13.25	-1	1.32	1	Technology
95	VXUS	11.17	-2	1.71	3	n/a
96	WDC	5.33	-1	0.53	1	Technology
97	WFM	53.08	-1	5.31	1	Consumer Services
98	WOOF	5.13	-8	0.54	2	Consumer Non-Durables
99	Z	43.88	-1	17.12	9	Miscellaneous
100	ZG	85.96	-9	18.36	4	Miscellaneous
101	ZION	31.92	-1	3.66	3	Finance