ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

Information theoretic causality detection between financial and sentiment data

Roberta Scaramozzino
(Università di Pavia)

Paola Cerchiello
(Università di Pavia)

Tomaso Aste
(University College London)

 # 202 (04-21)

Via San Felice, 5
I-27100 Pavia

economiaweb.unipv.it

Article

Information theoretic causality detection between financial and sentiment data

Roberta Scaramozzino [1]*, Paola Cerchiello [1] and Tomaso Aste [2]

[1] University of Pavia, Via San Felice 7, 27100 Pavia, Italy;
[2] University College of London, Gower Street, WC1E 6EA, London, United Kingdom;
* Correspondence: roberta.scaramozzino01@universitadipavia.it;

Abstract: The interaction between the flow of sentiment expressed on blogs and media and the dynamics of the stock market prices are analyzed through an information-theoretic measure, the transfer entropy, to quantify causality relations. We analyzed daily stock price and daily social media sentiment for the top 50 companies in the S&P index during the period from November 2018 to November 2020. We also analyzed news mentioning these companies during the same period. We found that there is a causal flux of information that links those companies. The largest fraction of significant causal links are between prices and between sentiments, but there is also significant causal information which goes both ways from sentiment to prices and from prices to sentiment. We observe that the strongest causal signal between sentiment and prices is associated with the Tech sector.

Keywords: Information theory; Textual analysis; Transfer Entropy; Financial news; Causality; Time Series

Citation: Scaramozzino, R.; Cerchiello, P.; Aste, T. Information theoretic causality detection between financial and sentiment data. *Entropy* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Causality is hard to detect from observations. This is because the occurrence of two events, one after the other, does not necessarily imply that the first caused the second. In the 1969 Granger [1] first proposed to look at causality in terms of the amount of extra information that the observation of a variable provides about another variable. In its original formulation this corresponds to an additional term in a liner regression for financial forecasting, but the idea is general and requires the quantification of information flow between variables.

In finance, the relationships between companies are usually analyzed considering the so-called "hard" information such as stock prices, trade volumes, the quantity of output but, in recent years, there has been an increase in the use of "soft" information including textual data, opinions, news and sentiment. Indeed, the economic value of things and firms is both material and immaterial. Reputation is playing a major role in economics. This has probably been always true, but it has become even more crucial in the present world where social-media have a pervasive role. Therefore, current study of market behaviour cannot be limited to the *hard* evidences related to the financial metrics but must also dig into the $soft$ metrics of social media and news. The relation between the two is still a domain in exploration. On one hand, efficient market hypothesis would suggest that all information must be comprised into the prices. On the other hand, swings in social opinions have their independent dynamics and sometime follow and other times anticipate market movements. In this paper, we further investigate such relationship by means of information theory tools, with the aim of understanding the manifest and latent dynamics of *hard* and $soft$ information within the US market.

We analyze the causality between some of the most important worldwide companies using both hard (prices) and soft (social media sentiment) information and investigate their interrelations. Causality is quantified through tools of information theory using entropy and mutual information. The first represents the uncertainty related to the variable's possible outcomes, the second one measures the information that two variables share [2].

1.1. Background: Textual analysis in finance

The use of textual analysis in the financial sector is relatively recent, but constantly growing.

Among the earlier papers, Engelberg [3] demonstrates that soft information, although more difficult to calculate, offers greater predictability on asset prices in particular at a longer horizon. Tirea and Negru [4] create an optimized portfolio through the combination of text mining, sentiment analysis, and risk models on the Bucharest Stock Exchange. Jothimani et al. [5] in their study integrate hard and soft data, the latter collected from online articles and tweets, and demonstrate that the combination of the two types of information allows optimization of the investment portfolio. Zheludev et al. [6] using sentiment techniques on social media messages show that, analyzing S&P index, information contained in social media can impact financial market forecasts.

With a focus on the impact of negative sentiment, Tetlock [7], using daily content from the Wall Street journal, finds that the volume of market exchanges is determined by unusually high or low pessimistic values. Indeed, Huang et al. [8] show that investors react differently depending on whether the information received is positive or negative; in the latter case the reaction is stronger. They also find, on a non-market-based test, evidence that information extracted from analyst reports has predictive power on earnings growth over the following 5 years.

Due to the easier processing of short text data, a notable application of sentiment analysis in finance has involved the analysis of tweets. Bollen et al. [9] examine whether the collective mood (based on 7 social moods), obtained from all the tweets published in a given period in USA, is correlated or predictive of DJIA values. They observe that only some of the 7 moods are correlated with DJIA values, with a lag of 3-4 days. Zhang et al. [10] find that, by analyzing the sentiment spikes on twetter posts, it is possible to predict what will happen in the market the following day. Rao et al. [11] using Granger's Causality Analysis show that, in the short term, tweets influence the trend in stock prices; Ranco et al. [12] considering 30 joint-stock companies of the Jones Industrial Average (DJIA) index, through the "study of events" methodology, they relate the prevailing sentiment in peak moments of tweets, in terms of volume, and stock returns showing a statistically significant dependence. Souza et al. [13] studying retail brands, analyze if there is significant connection between sentiment and volume of tweets with volatility and return on stock prices, seeing that the data obtained from social media are relevant to understand the financial dynamics and in particular, demonstrate how the sentiment obtained from the tweets is linked to the returns more than traditional news-wires.

You and Luo [14] investigate classification accuracy using textual and visual data. Carvalho et al. [15] classify tweets through an approach where paradigm words are selected using a genetic algorithm.

Kolchyna et al. [16] describe different techniques for classification of Twitter messages: lexicon based method and machine learning method, and present a new method that combine the two techniques. The score obtained from lexicon based method is the input feature for the machine learning approach and they demonstrate that classifications are more accurate using this combined technique.

In the field of financial risk management, Cerchiello and Giudici [17] construct a systemic risk model with a combination of financial tweet and financial price to comprehensively assess the impact of systemic risk.

1.2. Background: Information theory

Information theory was born in 1948 with the publication of Claude Shannon's article [18]. It stands at the interface of several multidisciplinary fields of research such as: mathematics, statistics, physics, telecommunications and computer science and it is applied to various fields, including the financial one.

Particularly used in the financial field is the concept of entropy. Dimpfl and Peter [19] analyzing through entropy the flow of information between CDS and the bond market, show that information flows in both directions with the importance of the CDS market increasing over time. Kwon and Yang [20] using entropy, examine the flow of information between composite stock indices and individual stocks and show that this flow is stronger from indices to stocks than vice versa. Shreiber [21] theorizes the concept of transfer entropy as a measure of coherence statistics between systems that evolve over time and Marschinski eand Kants [22], following this concept, analyze the flow of information between two time series: Dow Jones and DAX stock index. They introduce a modified estimator able to perform well also in case of short temporal series. Baek et al. [23] analyze, in the US stock market, the strength and direction of information using Transfer Entropy and conclude that companies in the energy and electricity sector influence the entire market. Nicola et al. [24] analyze the US banking network, made up of the top 74 listed banks, with the aim of highlighting whether mutual information and transfer entropy are capable of Granger cause financial stress indices and the USD / CHF exchange rate. For the implementation of the analysis they used general and partial granger causality, the latter correlated to representative measures of the general economic condition.

The main goal, in the present work, is to investigate the causal relationship between two events. We chose the asymmetric information-theoretic measure identified as transfer entropy, to detect strength and direction of transfer information between sentiment and prices, taking the advantage of application in the non-linear case differently from Granger Causality.

The design of the paper is organized as follows: Section 2 presents the methodology used, Section 3 presents a description of the data, in Section 4 we report the results and Conclusion are presented in Section 5.

2. Methods

In our work, we use a non-linear transfer entropy estimation, first introduced in [25], to identify and quantify causality between time series.

Using Shannon's measure of information [18], we can denote the uncertainty associated with a variable X by:

$$H(X) = -\sum_{x} p(x) \log p(x), \tag{1}$$

this quantity can be conditioned on a second variable to obtain conditional entropy:

$$H(X|Y) = H(X,Y) - H(X); \tag{2}$$

while the information that X and Y share is instead the so-called mutual information:

$$I(X,Y) = H(Y) - H(Y|X). \tag{3}$$

It expresses how the knowledge of a variable reduces the uncertainty of another and it is symmetric in X and Y.

We can express the information transfer from X to Y in terms of conditional mutual information for a given lag k:

$$TE^{(k)}_{(X\to Y)} = I(Y_t, X_{t-k}) = H(Y_t|Y_{t-k}) - H(Y_t|X_{t-k}, Y_{t-k}). \tag{4}$$

Eq.4 quantifies the amount or uncertainty on Y_t reduced by the knowledge of the lagged variable X_{t-k} given the information of the lagged variable Y_{t-k} itself. It is therefore a quantification of the additional information on variable Y provided by the past of variable X taking into account for what is already known about the past of Y.

This expression is general and applies to either linear and non linear estimations. In the liner case, one uses multivariate normal modeling, in the non–linear case one can instead estimate Transfer Entropy with a non-parametric density estimation which uses directly the empirical frequencies of observations into histogram bins.

In this paper, following [25], we adopt such non-parametric, non-linear approach and estimate the joint entropy using the multidimensional histogram tool available from the 'PyCausality' Python package [1]. According to such method, the observation space is divided into bins and the observations are allocated to each bin depending on their value. It is evident that the appropriate choice of Bins is crucial. We chose the equi-probable bins approach, which enforces that in each bin the number of data points is approximately the same. In previous studies [25], it was shown that this approach yields to best results for artificial data where the true underlying causality structure is known. In our case, where the causality structure must be discovered, we verified that other choices, such as equi-sized bins return similar results on our dataset, however the equi-probable bins provides cleanest outputs.

Another important choice is the lag k. We chose the first-order lag $k = 1$, since we assume that one day of delay is enough to see the effects of a variable on another. This is because, in an increasingly connected world, news spread almost immediately around the world. Similarly, the time for one event to impact another is extremely close.

The transfer entropy returns a non-negative real value. The greater the number, the larger is the amount of information measured. However, there is no reference and the number itself, without a benchmark, is of little interest. In order to obtain such a reference, we compared it with a null-hypothesis from data sets where any causal relation is removed. Such data were obtained from the original ones by shuffling randomly the time sequence of observations. In this way we obtained both a null-hypothesis reference and its statistics. From the mean $\left\langle TE_{shuffle} \right\rangle$ and the standard deviation $\sigma_{shuffle}$ of the shuffled transfer entropy we computed the significance of the Transfer entropy results in terms of the following Z-score:

$$Z := \frac{TE - \left\langle TE_{shuffle} \right\rangle}{\sigma_{shuffle}}. \tag{5}$$

The Z-score provides a distance, measured in terms of standard deviations, of the observed transfer entropy with respect to expected value for non-causally related variables. Larger Z-scores imply a value of the transfer entropy that is more significantly deviating from the values expected when the variables are not causally related, implying therefore a larger likelihood of causal relation. In this paper we used 50 shuffles.

Finally, we made use of the Z-score to construct graphs of significant causal links by retaining causality links at different threshold values, namely $Z > 2$ and $Z > 3$. The resulting networks were further considered in terms of community detection algorithms to identify causality structures. We also compared the networks between themselves and with respect to a reference network based on news.

3. Data

In this paper we consider the top 50 companies of S&P. The complete list of companies with the corresponding ticker code and rank Capitalization is available in Table 1.

[1] https://github.com/ZacKeskin/PyCausality

Table 1. Detailed description of the top 50 S&P companies.

Rank	Stock	Ticker
	Communication	
13	AT & T Inc.	T
18	Verizon Comm. Inc.	VZ
	Consumer Discretionary	
3	Amazon.com Inc.	AMZN
26	Comcast Corp.	CMCSA
14	Walt Disney Co.	DIS
19	Home Depot Inc.	HD
34	McDonald's Corp.	MCD
37	Netflix Inc.	NFLX
	Consumer Staples	
39	Costco Wholesale Corp.	COST
24	Coca-Cola Co.	KO
28	PepsiCo Inc.	PEP
10	Procter & Gamble Co.	PG
43	Philip Morris Int. Inc.	PM
30	Walmart Inc.	WMT
	Financial	
12	Bank of America Corp	BAC
5	Berkshire Hathaway Inc.	BRK.B
29	Citigroup Inc.	C
6	JPMorgan Chase & Co.	JPM
22	Wells Fargo & Co.	WFC
	Industrial	
25	Boeing Co.	BA
42	Honeywell Int. Inc.	HON
47	Union Pacific Corp.	UNP
50	Raytheon Technologies	RTX
	Healthcare	
41	AbbVie Inc.	ABBV
31	Abbott Laboratories	ABT
36	Amgen Inc.	AMGN
38	Bristol-Myers Squibb Co.	BMY
8	Johnson & Johnson	JNJ
33	Medtronic Plc	MDT
20	Merck & Co. Inc.	MRK
23	Pfizer Inc.	PFE
46	Thermo Fisher Scientific Inc.	TMO
15	UnitedHealth Group Inc.	UNH
	Tech	
2	Apple Inc.	AAPL
44	Accenture Plc	ACN
32	Adobe Inc.	ADBE
45	Broadcom Inc.	AVGO
35	Salesforce.com inc.	CRM
27	Cisco Systems Inc.	CSCO
4	Facebook Inc.	FB
7	Alphabet Inc.	GOOGL
16	Intel Corp.	INTC
17	Mastercard Inc.	MA
1	Microsoft Corp.	MSFT
40	NVIDIA Corp.	NVDA
49	Oracle Corp.	ORCL
48	PayPal Holdings Inc.	PYPL
9	Visa Inc.	V
	Energy	
21	Chevron Corp.	CVX
11	Exxon Mobil Corp.	XOM

We analyze two different types of information: stock prices and sentiment index.

The sentiment index is provided by Brain[2]. For each day, in a period starting from November 2018 to November 2020, a sentiment value is calculated from news and blog written in English. Brain sentiment indicator is represented by a value ranging between -1 to 1, where -1 corresponds to a negative sentiment, 0 to a neutral sentiment and + 1 to a positive sentiment.

For the same period, we have daily stock prices for each company from Yahoo finance. Since the sentiment index is available every day differently from market data, we exclude weekend days with regards to the former, so to have comparable time series.

For the prices detaset, we calculate the logarithmic return

$$L = \log(Price_t) - \log(Price_{t-1}), \tag{6}$$

which is a rate of change of the variable. We apply such transformation just to financial data because the sentiment index is already a stable variable in a range between -1 and 1. We performed the Anderson-Darling test and verified that all sentiment variables can be considered stationary with null-hypothesis p-values all below 5%.

After these pre-processing steps, we obtain a complete dataset, with values on the same scale for a total of 100 variables (50 prices log-returns and 50 sentiment index) and 514 observations (2 years daily data).

4. Results

As explained in the previous sections we want to assess the possible causal relationship between stock price and sentiment indicator focusing on some of the largest

[2] link to the site: https://braincompany.co/

worldwide companies. To this end, We compute the transfer entropy and the relative Z-score for all couples of variables (market price and sentiment index). We have therefore 100 variables and $100 \times 99 = 9,900$ distinct couples.

The full network of causality links without imposing any restriction is too dense. The large number of links and the significant density of the graph prevent from inferring useful and insightful information. A more detailed and consistent analysis is depicted in Figure 1 where it is shown a sub-network which retains only causal links with Z-scores larger than 3. Such a stringent score allows for the presence of the most significant links. Figure 1 clearly zoom on a fraction of the connections easing the readability. In this figure, and in all others, the clockwise direction of the arcs between nodes indicates the direction of connections. For a more comprehensive understanding, we report in Table 2 and Table 3 the associated Transfer Entropy values and the Z-score for each couple of stock with Z-score larger than 3.

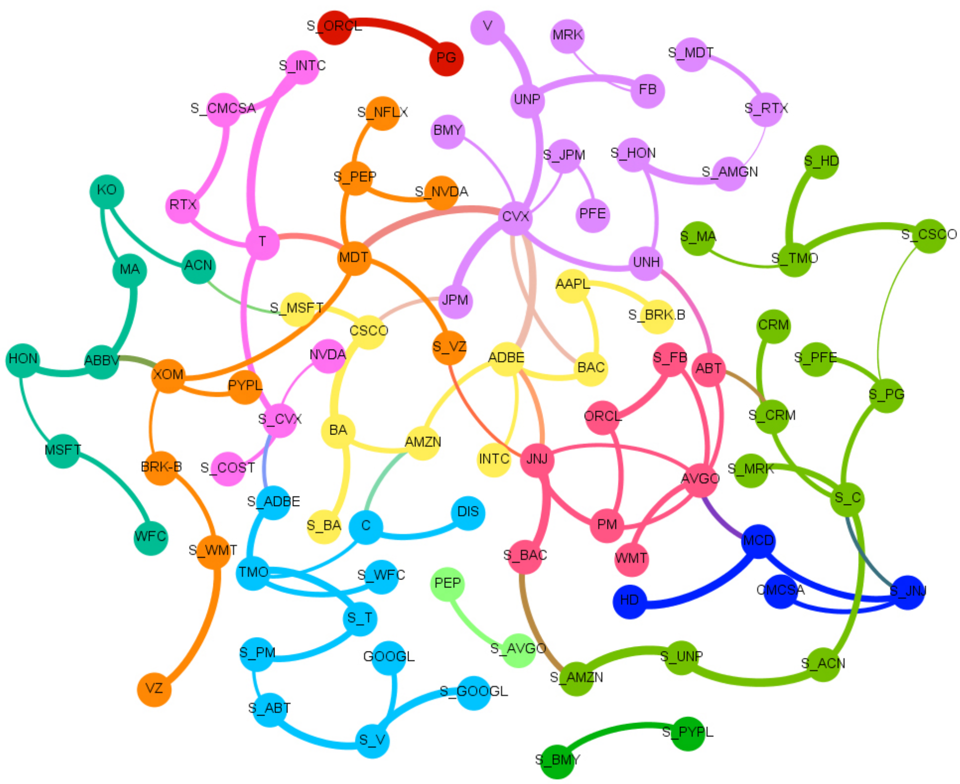

Figure 1. Network of links with Z score larger then 3. The colors represent the 12 Communities found using a Community detection algorithm. The sentiment index timeseries are indicated with an S before tickers name. The clockwise direction of the curves indicates the direction of connections.

Table 2. Couples of stocks with relative transfer Entropy, $TE^{(1)}_{(X \to Y)}$, values, Z scores larger than 3 (in brackets) and sectors for Price to Price network. The sectors are indicated with the capital letter, in particular we have F for Financial, H for Healthcare, T for Tech, I for Industrial, CD for Consumer discretionary, CS for Consumer staples, C for Communications, E for Energy.

Var X	Var Y	Value TE (Zscore)	Sectors
	Price to Price		
T	MDT	0.18 (4.24)	C→H
MSFT	WFC	0.18 (4.24)	T→F
PM	JNJ	0.18 (3.99)	CS→H
T	RTX	0.18 (3.98)	C→I
V	UNP	0.20 (3.98)	T→I
ABBV	HON	0.19 (3.81)	H→I
MCD	HD	0.19 (3.80)	CD→CD
MDT	CVX	0.19 (3.76)	H→E
UNP	FB	0.19 (3.75)	I→T
MSFT	HON	0.17 (3.66)	T→I
WMT	AVGO	0.18 (3.64)	CS→T
BAC	ADBE	0.18 (3.64)	F→T
JPM	CVX	0.20 (3.63)	F→E
UNP	CVX	0.19 (3.61)	I→E
ABBV	XOM	0.18 (3.54)	H→E
DIS	C	0.18 (3.38)	CD→F
MA	ABBV	0.19 (3.3)	T→H
C	AMZN	0.18 (3.36)	F→CD
AVGO	PM	0.1 (3.35)	T→CS
BA	CSCO	0.2 (3.35)	I→T
AAPL	BAC	0.18 (3.34)	T→F
UNH	ABT	0.18 (3.33)	H→H
CVX	ADBE	0.19 (3.33)	E→T
BRK-B	XOM	0.17 (3.26)	F→E
ORCL	PM	0.18 (3.24)	T→CS
MA	KO	0.18 (3.24)	T→CS
ADBE	INTC	0.18 (3.24)	T→T
BAC	CVX	0.18 (3.22)	F→E
ADBE	JNJ	0.18 (3.22)	T→H
C	TMO	0.18 (3.16)	F→H
FB	MRK	0.17 (3.15)	T→H
AMZN	BA	0.18 (3.13)	CD→I
MDT	XOM	0.18 (3.11)	H→E
BMY	CVX	0.17 (3.11)	H→E
PYPL	XOM	0.18 (3.10)	T→E
CSCO	JPM	0.18 (3.1)	T→F
UNH	CVX	0.19 (3.06)	H→E
ABT	AVGO	0.18 (3.05)	H→T
ACN	KO	0.18 (3.04)	T→CS
JNJ	AVGO	0.18 (3.04)	H→T
AMZN	ADBE	0.18 (3.02)	CD→T
MCD	AVGO	0.18 (3.00)	CD→T

Table 3. Couples of stocks with relative transfer Entropy, $TE^{(1)}_{(X \to Y)}$, values, Z scores larger than 3 (in brackets) and sectors for Price to Sentiment, Sentiment to Sentiment and Sentiment to Price networks. The sectors are indicated with the capital letter, in particular we have F for Financial, H for Healthcare, T for Tech, I for Industrial, CD for Consumer discretionary, CS for Consumer staples, C for communications, E for Energy.

Var X	**Var Y**	**Value TE (Zscore)**	**Sectors**
	Sentiment to Sentiment		
AMGN	HON	0.2 (4.63)	H→I
AMZN	UNP	0.2 (4.57)	CD→I
C	CRM	0.18 (4.51)	F→T
C	ACN	0.19 (4.47)	F→T
TMO	CSCO	0.19 (4.36)	H→T
AMZN	BAC	0.19 (4.34)	CD→F
BMY	PYPL	0.19 (4.3)	H→T
TMO	HD	0.2 (4.04)	H→CD
V	ABT	0.2 (3.97)	T→H
V	GOOGL	0.2 (3.89)	T→T
INTC	CMCSA	0.19 (3.82)	T→CD
ACN	UNP	0.2 (3.79)	T→I
NVDA	PEP	0.18 (3.57)	T→CS
MRK	C	0.19 (3.45)	H→F
T	PM	0.19 (3.42)	C→CS
PFE	PG	0.18 (3.33)	H→CS
ABT	PM	0.17 (3.32)	H→CS
TMO	MA	0.17 (3.32)	H→T
C	PG	0.18 (3.31)	F→CS
MDT	RTX	0.18 (3.12)	H→I
CVX	COST	0.18 (3.09)	E→CS
PEP	NFLX	0.18 (3.08)	CS→CD
JNJ	C	0.18 (3.07)	H→F
ADBE	CVX	0.18 (3.07)	T→E
RTX	AMGN	0.16 (3.04)	I→H
PG	CSCO	0.16 (3.03)	CS→T

Var X	**Var Y**	**Value TE (Zscore)**	**Sectors**
	Sentiment to Price		
CVX	T	0.19 (4.34)	E→C
ORCL	PG	0.20 (4.24)	T→CS
FB	ORCL	0.19 (4.01)	T→T
WMT	VZ	0.12 (3.83)	CS→C
WFC	TMO	0.18 (3.68)	F→H
MSFT	ACN	0.17 (3.64)	T→T
CMCSA	RTX	0.19 (3.61)	CD→I
JNJ	CMCSA	0.18 (3.41)	H→CD
AVGO	PEP	0.18 (3.38)	T→CS
JNJ	MCD	0.19 (3.37)	H→CD
JPM	PFE	0.17 (3.29)	F→H
HON	UNH	0.18 (3.19)	I→H
CVX	NVDA	0.17 (3.17)	E→T
MSFT	CSCO	0.19 (3.12)	T→T
JPM	CVX	0.17 (3.06)	F→E
CRM	CRM	0.19 (3.06)	T→T
FB	AVGO	0.18 (3.03)	T→T
	Price to Sentiment		
JNJ	BAC	0.2 (4.37)	H→F
TMO	ADBE	0.19 (4.05)	H→T
TMO	T	0.19 (3.92)	H→C
T	INTC	0.20 (3.83)	C→ T
ABT	CRM	0.18 (3.54)	H→T
BA	BA	0.19 (3.51)	I→I
MDT	VZ	0.18 (3.36)	H→C
AAPL	BRK.B	0.18 (3.34)	T→F
BRK-B	WMT	0.18 (3.18)	F→CS
JNJ	VZ	0.17 (3.08)	H→C
GOOGL	V	0.18 (3.06)	T→T
MDT	PEP	0.18 (3.05)	H→CS

The two tables report results classified according to the S&P industry sectors: Consumer discretionary, Consumer staples, Energy, Healthcare, Tech, Financial, Industrial and Communications. The sectors are not homogeneously populated, in particular, Healthcare and Tech ones have the largest number of stocks, respectively, 10 and 15 companies. Whilst the sectors classification is important for the correct assessment of the pattern drivers, it is unquestionable the tendency of big companies to diversify more and more the types of business. As an example, Amazon, which is listed in the Consumer discretionary sector, has a division named 'Amazon Web Services' for cloud computing and device and a division named 'Amazon Studios' for music and videos streaming. This to bear in mind that the division among the sectors does not completely reflect the real connections among the companies.

A Community Detection algorithm [26] is employed to investigate the presence of meaningful communities inside our network in Figure 1. The community algorithm finds 12 different communities as we can see from the different colors. Most of the communities are similar in terms of number of companies. Interestingly, such groups have some recognizable overlap with S&P sectors, but also distinctive features revealing the different nature of market price and sentiment interconnections which goes well beyond companies core business.

By looking at the connections in such a network we can distinguish between variables associated to the price returns (identified generically as 'price' hereafter) and variables instead associated with sentiment scores (identified generically as 'sentiment' hereafter).

We observe that the most of the links are from Price to Price (See Table 2), followed by the links from Sentiment to Sentiment and then the Sentiment to Price and finally Price to Sentiment (see Table 3). We observe an interesting asymmetry between companies and sectors that are influencers and the others that are followers with most of the significant links involving two different industry sectors. The leading one, in terms of number of significant links, is the Technological sector with a predominance of connection towards the Consumer sector: Accenture causing ($\rightarrow$) Coca-Cola; Mastercard $\rightarrow$ Coca-Cola; Broadcom $\rightarrow$ Philip Morris; Oracle $\rightarrow$ Philip Morris; Amazon $\rightarrow$ Adobe; McDonald's $\rightarrow$ Broadcom; Walmart $\rightarrow$ Broadcom. Very interesting is also the influence of different sectors onto the Energy one: Bank of America, Bristol, JPMorgan, Medtronic, UnitedHealth and Union Pacific cause Chevron; while Paypal causes Exxon. We note that this aboundance of links to the energy sector is unique to this Price to Price network. Within the same sector. There are also several links within the same sectors: a connection between United health $\rightarrow$ Abbot, both in the Healthcare sector; McDonald's $\rightarrow$ Home Depot, in the Consumer sector; and Adobe $\rightarrow$ Intel in the Tech sector.

There are also, numerous links in the Sentiment to Sentiment network (see in Table 3). In this case, many links are related to the Healthcare sector, most of them are relationships between the Healthcare and the Consumer sector: Johnson&Johnson $\rightarrow$ Walt Disney; Merck&Co $\rightarrow$ Walt Disney; Thermo Fisher $\rightarrow$ Home Depot; Pfizer $\rightarrow$ Procter&Gabmble; Abbott $\rightarrow$ Philip Morris. We find also links between companies in the same sector: Pepsi $\rightarrow$ Netflix; and Walt Disney $\rightarrow$ Procter&Gamble.

In the Price to Sentiment network (Table 3), we notice that there is a significant frequency of stocks related to the Healthcare sector which affect other sectors: Tech (Thermo Fisher $\rightarrow$ Adobe, Abbott $\rightarrow$ Salesforce.com); Financial (Johnson&Johnson $\rightarrow$ Bank of America); Consumer (Medtronic $\rightarrow$ Pepsi); and Communications (Thermo Fisher $\rightarrow$ AT&T, Johnson&johnson $\rightarrow$ Verizon and Medtronic $\rightarrow$ Verizon).

Perhaps, the most interesting result lays upon the causal links from Sentiment to Price. Most of them are in the Technological sector in particular Tech to Tech: Microsoft $\rightarrow$ Accenture; Facebook $\rightarrow$ Broadcom; Salesforce.com, Microsoft $\rightarrow$ Cisco; and Facebook $\rightarrow$ Oracle.

The analysis reveals a dominant role of Healthcare and Technology both as influencer and follower sectors across all four networks. Another important sector is Consumer, both essential (staples) and discretionary, which are however mainly followers and less influencers.

To ease the interpretation, we report in Figures 2, 3, 4 and 5 an aggregated network visualization of Tables 2 and 3 representing the flows of influence between industry sectors quantified as total, significant ($Z > 3$), transfer entropy exchanged in each direction. This analysis allows for a global view of the 8 sectors in terms of reciprocal influence. We note that the four networks have very distinct characteristics.

Specifically, in the Price$\rightarrow$Price network in Figure 2 we observe a role of the energy sector, being a follower of both Financial and Healthcare sectors; a role that is not revealed in any of the other networks. Moreover we stress that the financial sector, which traditionally plays a pivotal role when the financial market is considered, appears to be not so predominant. Indeed, the largest average Transfer Entropy is measured from Healthcare to Energy with 0.92.

The Sentiment$\rightarrow$Price network in Figure 3 has a mayor self-influencing loop with the sentiment on the Technological sector affecting its own price (TE 0.92); it also reveals some influence of the Financial sector on Healthcare (TE 0.36) and Healthcare on Consumer Discretionary (TE 0.37).

In the Price$\rightarrow$Sentiment network in Figure 4 the main leading role is played by Healthcare and it also emerges role of the Communication sector as follower of Healthcare (TE 0.55) and as influencer of Technology (TE 0.2). This is not present in any of the other networks. Healthcare is also influencing Technology (TE 0.37).

Finally, the Sentiment→Sentiment network in Figure 5 shows a dominating role of Healthcare which is affecting the Consumer sectors (TE 0.56), Industry (TE 0.38) and Technology (TE 0.55).

Overall, the Pirce→Price network has the largest number of connections i.e. 25, then Sentiment→Sentiment follows with 19, finally Sentiment→Price and Pirce→Sentiment with respectively 10 and 9.

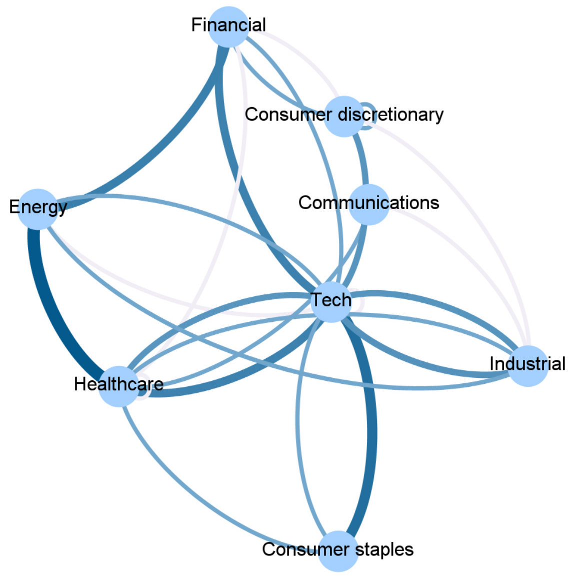

Figure 2. The aggregated Price → Price network visualization of Tables 2 and 3 representing the flows of influence among sectors quantified as total, significant (Z>3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

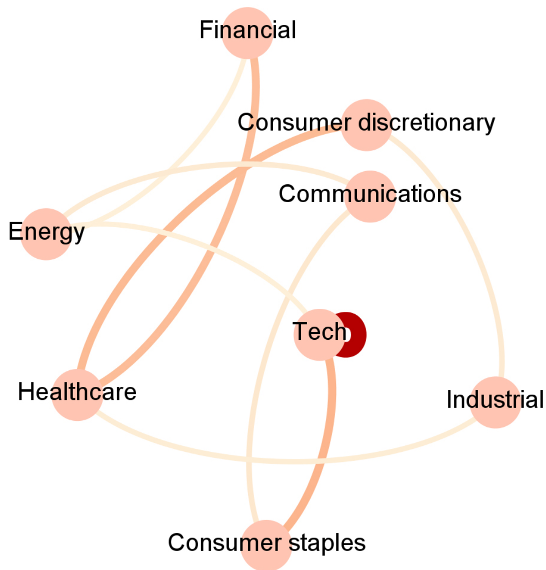

Figure 3. The aggregated Sentiment → Price network visualization of Tables 2 and 3 representing the flows of influence among sectors quantified as total, significant (Z>3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

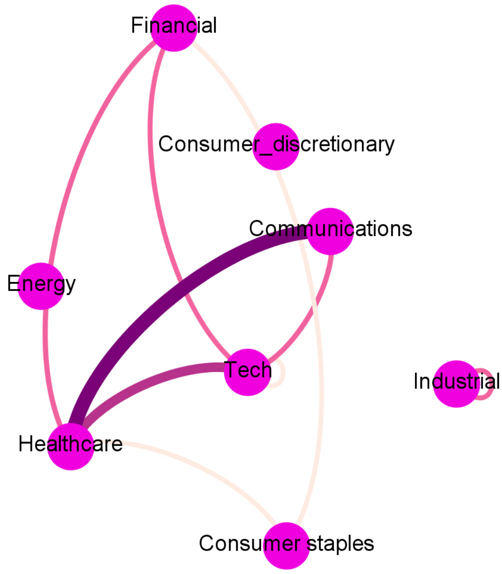

Figure 4. The aggregated Price → Sentiment network visualization of Tables 2 and 3 representing the flows of influence among sectors quantified as total, significant (Z>3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

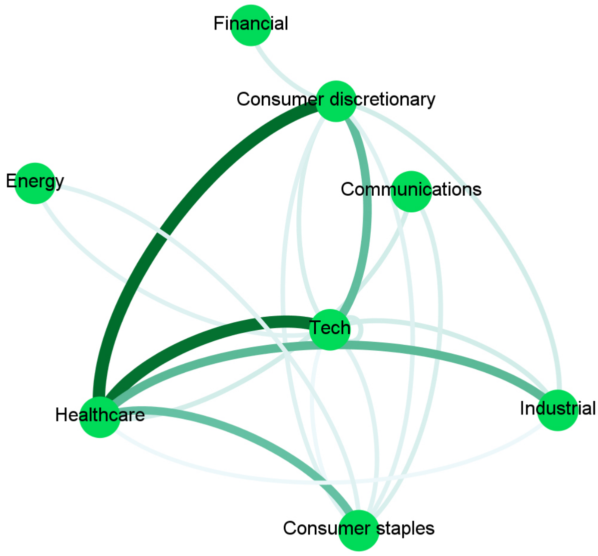

Figure 5. The aggregated Sentiment → Sentiment network visualization of Tables 2 and 3 representing the flows of influence among sectors quantified as total, significant (Z>3), transfer entropy exchanged in each direction. The clockwise direction of the curves indicates the direction of connections.

4.1. Comparison between TE matrix and dataset based on News

Since one of the main aim of our paper is to disentangle the role played by the information disclosed through news and measured by means of a sentiment score we further analyze such component. To deepen our investigation we pay greater attention to the sentiment aspect carring out a further analysis using data concerning news provided by Brain,[3] to identify relations between stocks by counting the number of times two tickers are mentioned within the same news article.

In Figure 6 we report the complete network of news in common. As already happened with unrestricted analysis, the network appears too dense to be readable. However some clear patterns are already evident, like the strict connections among the company giants like AAPL, MSFT, GOOGL, FB, AMZN (bottom right in blue) which indeed represent a community per se.

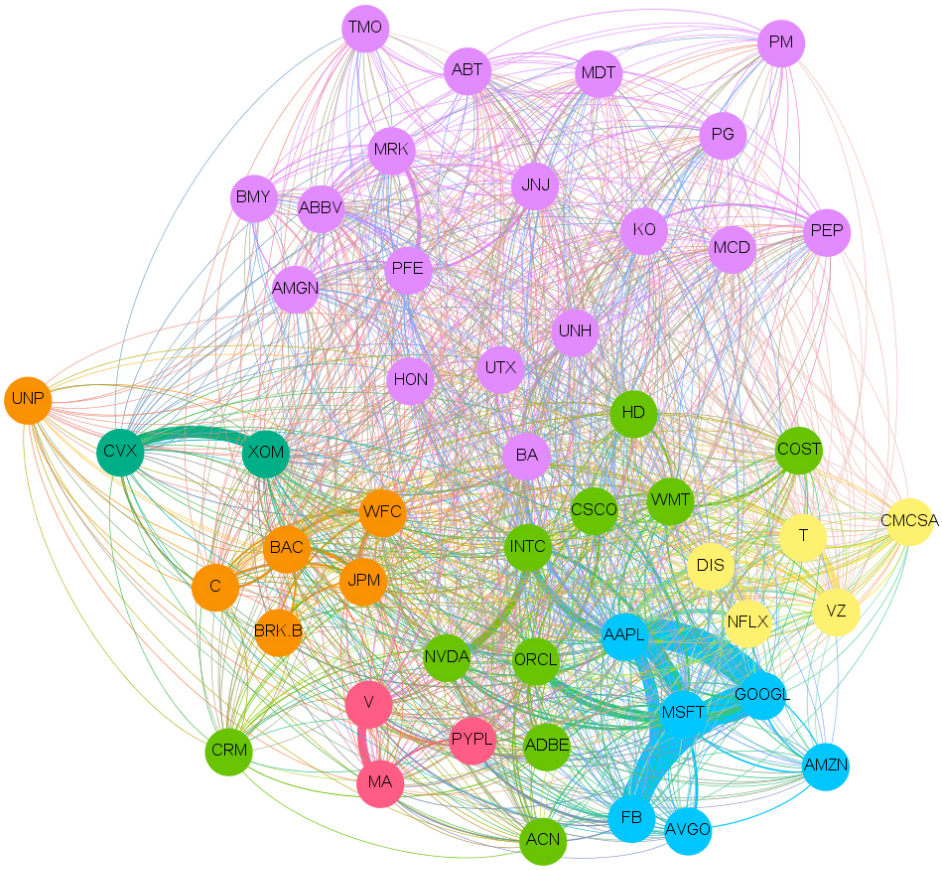

Figure 6. Network news in common. The colours represent the 7 Communities found using a Community detection algorithm. The clockwise direction of the curves indicates the direction of connections.

To ease the readability we filter out the less significant links, thus in Figure 7 we report the network built by retaining only the connections between stocks that score a number of news in common larger than a threshold value of 20 (such value has been identified after some sensitivity analysis).

[3] link to the site: https://braincompany.co/

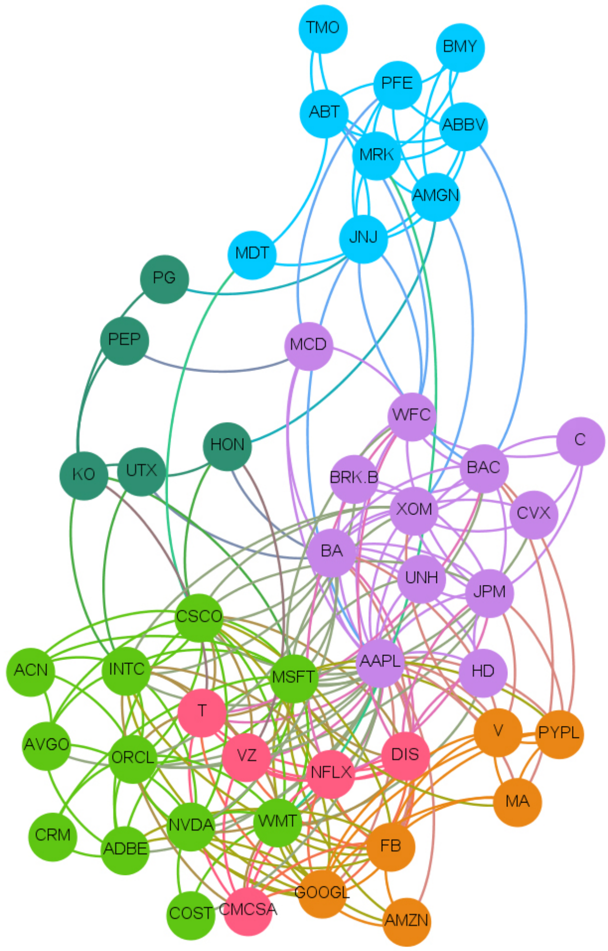

Figure 7. Network news in common larger than 20. The colours represent the 7 Communities found using a Community detection algorithm. The clockwise direction of the curves indicates the direction of connections.

Such a network is then compared with the previous causality networks for Price to Price (PP) figure 2, Sentiment to Price (SP) figure 3, Price to Sentiment (PS) figure 4 and Sentiment to Sentiment (SS) figure 5 obtained by imposing on the links a threshold Z-score value.

Results for the thresholds: $Z > 2.5$ and a number of news in common larger than 20 are reported in Table 4. The reader can see that there is a rather modest overlap between the networks that mostly involves very popular companies.

In order to statistically quantify the significance of such overlap between the networks, we compute the hypergeometric probability to have a certain number or more of overlaping edges in two directed graphs. Of course results depend upon the chosen thresholding for the number of news and the Z-score. Overall we find that there is no statistical significance in terms of p-value for the thresholds $Z > 2.5$ and News > 20. However, this does not mean that the links are just by chance.

By performing a sensitivity analysis by changing the threshold values, we observe that, the 4 networks have different patterns. The Price to Price causality network shows relations with news with a rather large number of overlaps and statistical significance with p-values below 1% but only when the network is less restricted using small news threshold and small Z-scores. This seems to indicate that news pick some insights of the internal dynamics of the market and that identify correctly important events in the financial domain which trigger propagation of information through the social media.

Table 4. Overlap between links in news network and links in Transfer Entropy matrix with a threshold on news equal to 20 and on Z-score equal to 2.5.

var_x	var_y
Price to Price variables (PP)	
NVDA	BA
BAC	AAPL
CMCSA	T
CSCO	BA
CSCO	NVDA
CSCO	ORCL
HD	JPM
INTC	T
JPM	CSCO
BA	NVDA
NVDA	MSFT
PYPL	JPM
PYPL	MSFT
WFC	MSFT
Price to Sentiment variables (PS)	
JNJ	BAC
ABBV	BMY
AAPL	BRK.B
BAC	BRKB.B
T	INTC
GOOGL	V
CSCO	WMT

var_x	var_y
Sentiment to Price variables (SP)	
MSFT	AAPL
MSFT	ACN
MSFT	CSCO
MSFT	GOOGL
CRM	ORCL
MSFT	PYPL
ABT	TMO
Sentiment to Sentiment variables (SS)	
FB	ADBE
INTC	CMCSA
ADBE	CRM
INTC	CSCO
V	GOOGL
AMGN	HON
FB	V
XOM	MSFT

This significance at small thresholds could indicate that this happens on average but the importance of the news or the intensity of the causality relation is not relevant.

For what concerns the other networks we observe that larger thresholds (more restrictive condition and less links) for the number of news in common increase statistical significance. This could indicate that news are identifying events that also resonate on the social media but this tend to happen only for events with high relevance.

5. Discussion and Conclusion

In this paper, we study the causal relationships between opinion reflected on blogs and media and the patterns in stock market values, to investigate causal interactions between these variables. We focus on top 50 companies of the S&P index rooted in different sectors: Consumer discretionary, Consumer staples, Energy, Healthcare, Tech, Financial Industrial and Communications. Data covers two years from November 2018 through November 2020. In our analysis we employ an information-theoretic measure, the transfer entropy, to monitor the information flows between sentiment and market movements. We use a recently developed non-linear methodology [25] that can better capture causality extending the traditional Granger approach.

Our information-theoretic analysis revealed a large number of strong connections. As expected, the highest number of significant causal relationships between companies involves the same kind of data source (price $\rightarrow$ price, sentiment $\rightarrow$ sentiment) but there are also strong connections cross-sources. Some sectors are more influential in terms of sentiment dynamics and less in terms of price dynamics. For instance, in the sentiment to sentiment network we can clearly spot the pivotal role of the Healthcare sector which influences both the consumer discretionary and the technological sectors. Such pattern is present, although with differentiated importance within the other networks too. What surprises is the role of the Financial sector which is traditionally in a paramount position compared to other sectors. Our analysis shows that financial companies are still important if we restrict to price data solely or if we consider the impact of sentiment on price but much less within the alternative scenarios. However, this is in line with what already reported in [27] were a reduction of centrality of the financial sector was pointed out. This was also reported by [28], where through a temporal dynamic network analysis the authors shows that the financial sector behaves differently as an isolated cluster which reacts mainly to market price data. Another important sector is the technological one, either as influencer or follower depending on the network we may consider. The remaining sectors seem less consistent and change in relevance and role across the different networks.

From this study we can conclude, first of all, that mutual influences between various companies are not limited to influences between companies within the same sector. On the contrary, the cross sector interactions tend to be more relevant. This might be because companies with high capitalization tend to operate in many markets other than their core business. Secondly, the price variables show a more homogeneous behavior, with connections which tend to be stronger and also more frequent. Nonetheless, we identify several cases where sentiment about a company has strong influence to sentiment on other companies and also to other company prices. In particular the Tech sector reveals a very strong influence of sentiment on prices. This might be a consequence of the presence of the most popular companies in terms of branding, the 'Big Five' (Google, Amazon, Facebook, Microsoft and Apple), which are often mentioned in news and blogs and this continuous notoriety obviously affects the financial aspect.

Author Contributions: Conceptualization, R.S. and T.A.; Formal analysis, R.S.; investigation, R.S. and T.A. and P.C.; data curation, R.S.; writing—original draft preparation, R.S.; writing—review and editing, T.A. and P.C.; supervision, T.A. and P.C. All authors have read and agreed to the published version of the manuscript.

Funding: TA acknowledges support from ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018-2 825215). RS, PC and TA acknowledge support from FIN-TECH H2020-ICT-2018-2 825215

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Brain and are available from the authors (https://braincompany.co/) with the permission of Brain.

Acknowledgments: The authors thanks Brain in particular Matteo Campellone and Francesco Cricchio for sharing their data for sentiment score.

Conflicts of Interest: The authors declare no conflict of interest.

387 ## Appendix A

Table A1. Aggregated network for the following influencing sectors: Tech, Communications, Consumer Discretionary and Consumer Staples.

Source	Target	P→P	S→S	S→P	P →S
Tech	Consumer staples	0.72	0.18	0.39	0
Tech	Healthcare	0.54	0	0	0
Tech	Financial	0.54	0	0	0.19
Tech	Industrial	0.37	0.20	0	0
Tech	Energy	0.18	0.18	0	0
Tech	Tech	0.18	0.20	0.92	0.18
Tech	Consumer discretionary	0	0.19	0	0
Tech	Communications	0	0	0	0
Communications	Healthcare	0.19	0.20	0	0
Communications	Industrial	0.18	0	0	0
Communications	Tech	0	0	0	0.20
Communications	Consumer staples	0	0.19	0	0
Communications	Communications	0	0	0	0
Communications	Consumer discretionary	0	0	0	0
Communications	Financial	0	0	0	0
Communications	Energy	0	0	0	0
Consumer discretionary	Tech	0.37	0.37	0	0
Consumer discretionary	Consumer discretionary	0.20	0	0	0
Consumer discretionary	Financial	0.19	0.19	0	0
Consumer discretionary	Industrial	0.18	0.20	0.19	0
Consumer discretionary	Consumer staples	0	0.18	0	0
Consumer discretionary	Communications	0	0	0	0
Consumer discretionary	Healthcare	0	0	0	0
Consumer discretionary	Energy	0	0	0	0
Consumer staples	Healthcare	0.19	0	0	0
Consumer staples	Tech	0.19	0.16	0	0
Consumer staples	Communications	0	0	0.20	0
Consumer staples	Consumer discretionary	0	0.18	0	0
Consumer staples	Consumer staples	0	0	0	0
Consumer staples	Financial	0	0	0	0
Consumer staples	Industrial	0	0	0	0
Consumer staples	Energy	0	0	0	0

Table A2. Aggregated network for the following influencing sectors: Financial, Healthcare, Industrial and Energy.

Source	Target	P→P	S→S	S→P	P →S
Financial	Energy	0.56	0	0.17	0
Financial	Tech	0.19	0	0	0
Financial	Consumer discretionary	0.18	0	0	0
Financial	Healthcare	0.18	0	0.36	0
Financial	Consumer staples	0	0	0	0.18
Financial	Communications	0	0	0	0
Financial	Financial	0	0	0	0
Financial	Industrial	0	0	0	0
Healthcare	Energy	0.92	0	0	0
Healthcare	Tech	0.36	0.55	0	0.37
Healthcare	Industrial	0.19	0.38	0	0
Healthcare	Healthcare	0.18	0	0	0
Healthcare	Consumer discretionary	0	0.56	0.37	0
Healthcare	Consumer staples	0	0.36	0	0.18
Healthcare	Communications	0	0	0	0.55
Healthcare	Financial	0	0	0	0.20
Industrial	Tech	0.39	0	0	0
Industrial	Energy	0.20	0	0	0
Industrial	Industrial	0	0	0	0.19
Industrial	Healthcare	0	0.16	0.18	0
Industrial	Communications	0	0	0	0
Industrial	Consumer discretionary	0	0	0	0
Industrial	Consumer staples	0	0	0	0
Industrial	Financial	0	0	0	0
Energy	Tech	0.19	0	0.17	0
Energy	Communications	0	0	0.19	0
Energy	Consumer staples	0	0.18	0	0
Energy	Consumer discretionary	0	0	0	0
Energy	Financial	0	0	0	0
Energy	Healthcare	0	0	0	0
Energy	Industrial	0	0	0	0
Energy	Energy	0	0	0	0

References

1. Granger, C.W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* **1969**, pp. 424–438.
2. Cover, T.M. *Elements of information theory*; John Wiley & Sons, 1999.
3. Engelberg, J. Costly information processing: Evidence from earnings announcements. AFA 2009 San Francisco meetings paper, 2008.
4. Tirea, M.; Negru, V. Investment portfolio optimization based on risk and trust management. 2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY). IEEE, 2013, pp. 369–374.
5. Jothimani, D.; Shankar, R.; Yadav, S.S. A big data analytical framework for portfolio optimization. *arXiv preprint arXiv:1811.07188* **2018**.
6. Zheludev, I.; Smith, R.; Aste, T. When can social media lead financial markets? *Scientific reports* **2014**, *4*, 4213.
7. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance* **2007**, *62*, 1139–1168.
8. Huang, A.H.; Zang, A.Y.; Zheng, R. Evidence on the information content of text in analyst reports. *The Accounting Review* **2014**, *89*, 2151–2180.
9. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *Journal of computational science* **2011**, *2*, 1–8.
10. Zhang, X.; Fuehres, H.; Gloor, P.A. Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences* **2011**, *26*, 55–62.
11. Rao, T.; Srivastava, S.; others. Analyzing stock market movements using twitter sentiment analysis **2012**.
12. Ranco, G.; Aleksovski, D.; Caldarelli, G.; Grčar, M.; Mozetič, I. The effects of Twitter sentiment on stock price returns. *PloS one* **2015**, *10*, e0138441.
13. Souza, T.T.P.; Kolchyna, O.; Treleaven, P.C.; Aste, T. Twitter sentiment analysis applied to finance: A case study in the retail industry. *arXiv preprint arXiv:1507.00784* **2015**.
14. You, Q.; Luo, J. Towards social imagematics: sentiment analysis in social multimedia. Proceedings of the thirteenth international workshop on multimedia data mining, 2013, pp. 1–8.
15. Carvalho, J.; Prado, A.; Plastino, A. A statistical and evolutionary approach to sentiment analysis. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE, 2014, Vol. 2, pp. 110–117.
16. Kolchyna, O.; Souza, T.T.; Treleaven, P.; Aste, T. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955* **2015**.
17. Cerchiello, P.; Giudici, P. Big data analysis for financial risk management. *Journal of Big Data* **2016**, *3*, 18.
18. Shannon, C.E. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423.
19. Dimpfl, T.; Peter, F.J. Using transfer entropy to measure information flows between financial markets. *Studies in Nonlinear Dynamics & Econometrics* **2013**, *17*, 85–102.
20. Kwon, O.; Yang, J.S. Information flow between composite stock index and individual stocks. *Physica A: Statistical Mechanics and its Applications* **2008**, *387*, 2851–2856.
21. Schreiber, T. Measuring information transfer. *Physical review letters* **2000**, *85*, 461.
22. Marschinski, R.; Kantz, H. Analysing the information flow between financial time series. *The European Physical Journal B-Condensed Matter and Complex Systems* **2002**, *30*, 275–281.
23. Baek, S.K.; Jung, W.S.; Kwon, O.; Moon, H.T. Transfer entropy analysis of the stock market. *arXiv preprint physics/0509014* **2005**.
24. Nicola, G.; Cerchiello, P.; Aste, T. Information network modeling for US banking systemic risk. *Entropy* **2020**, *22*, 1331.
25. Keskin, Z.; Aste, T. Information-theoretic measures for non-linear causality detection: application to social media sentiment and cryptocurrency prices. *arXiv preprint arXiv:1906.05740* **2019**.
26. Fortunato, S. Community detection in graphs. *Physics reports* **2010**, *486*, 75–174.
27. Aste, T.; Shaw, W.; Di Matteo, T. Correlation structure and dynamics in volatile markets. *New Journal of Physics* **2010**, *12*, 085009.
28. Ahelegbey, D.F.; Cerchiello, P.; Scaramozzino, R. Network Based Evidence of the Financial Impact of COVID-19 Pandemic, 2021. [Online; accessed 30. Mar. 2021], doi:10.2139/ssrn.3780954.