Multimed Tools Appl
https://doi.org/10.1007/s11042-018-6388-4

Twitter permeability to financial events: an experiment towards a model for sensing irregularities

Ana Fernández Vilas[1] · Rebeca P. Díaz Redondo[1] · Keeley Crockett[2] · Majdi Owda[2] · Lewis Evans[2]

Received: 1 December 2017 / Revised: 22 May 2018 / Accepted: 5 July 2018

Abstract There is a general consensus of the good sensing and novelty characteristics of Twitter as an information media for the complex financial market. This paper investigates the permeability of Twittersphere, the total universe of Twitter users and their habits, towards relevant events in the financial market. Analysis shows that a general purpose social media is permeable to financial-specific events and establishes Twitter as a relevant feeder for taking decisions regarding the financial market and event fraudulent activities in that market. However, the provenance of contributions, their different levels of credibility and quality and even the purpose or intention behind them should to be considered and carefully contemplated if Twitter is used as a single source for decision taking. With the overall aim of this research, to deploy an architecture for real-time monitoring of irregularities in the financial market, this paper conducts a series of experiments on the level of permeability and the permeable features of Twitter in the event of one of these irregularities. To be precise, Twitter data is collected concerning an event comprising of a specific financial action on the 27th January 2017: the announcement about the merge of two companies Tesco PLC and Booker Group PLC, listed in the main market of the London Stock Exchange

✉ Ana Fernández Vilas
avilas@det.uvigo.es

Rebeca P. Díaz Redondo
rebeca@det.uvigo.es

Keeley Crockett
k.crockett@mmu.ac.uk

Majdi Owda
m.owda@mmu.ac.uk

Lewis Evans
L.Evans@mmu.ac.uk

[1] I&C Laboratory AtlantTIC Research Centre, University of Vigo, Vigo, 36310, Spain

[2] School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University, Manchester, M1 5GD, UK

Published online: 20 August 2018

(LSE), to create the UK's Leading Food Business. The experiment attempts to answer two research questions which aim to characterize the features of Twitter permeability to the financial market. The experimental results confirm that a far-impacting financial event, such as the merger considered, caused apparent disturbances in all the features considered, that is, information volume, content and sentiment as well as geographical provenance. Analysis shows that although Twitter is not a specific financial forum, it is permeable to financial events. Therefore it should be considered within the architecture for real-time monitoring of irregularities in the financial market.

Keywords Twitter data analysis · Stock market · Irregularity behaviour · London stock exchange

1 Introduction

Progressive usages of technology in the stock markets have led to a continuous growth in their business. Businesses and individual investors alike can take decisions and invest within minutes if not in seconds [24]. Helping both businesses and individual investors harvest information from diverse sources such as the company itself, through their website or Regulatory News Service (RNS), news agencies, brokers, stock market and individual investors, is a difficult and time consuming task. At the beginning of the emergence of social media, stock discussion forums pioneered an alternative source of information for investors, specially retailers, supplementing the traditional news media. A plethora of academic works study the capability of these highly-specialized social channels to predict returns and to detect abnormal behaviours in the stock market [3, 9, 15, 47]. From the investor's perspective, success highly depends on the quality and the rapidness of the information to support the decision-making. As online social media invaded the habits of people, also companies, brokers and other key roles in the financial market began to share more and more factual information and informed opinion in social media. The task of a financial analysts becomes progressively more challenging if we consider that not only have they to take decisions in seconds, but also deal with an increasing number of data sources comprising of a continuous range of different features and quality. Although social media may be a vehicle to fight against asymmetric information in the financial market [5], the volume, velocity and variety of its data make investing a heroic task, especially for individual investors.

Given today co-existence of social media and traditional news media (newspaper or online news media), the authors in [25] compare the performance of social media and news media in predicting returns in the Australian market. It is clear that social media outperforms traditional media on the capability for readers to interact or further spread information. Although the study shows a significant effect of sentiments on one online discussion forum focused on the Australian financial market, it found that the effect of new media was not significant. The most interesting finding was that sentiments on online investor forums do not seem to mirror the ones from news media.There is in fact, an ongoing debate about to what extent social media analytics would be the new predictor for the future returns of stock-exchange traded financial assets. We may even wonder if social media can lead the direction of the financial market at some point in the future as some authors suggested [59].

Although organizations have always used various corporate disclosure channels to communicate directly and indirectly with investors, they are now relying on social media to provide up-to-date information to investors. As the heart of publicly accessible Social Media, Twitter has become a vital source for open source social intelligence about news,

natural disasters, terrorist attacks, public health, politics, etc. Also, Twitter is one of the most currently used platforms to share financial information from companies, brokers, news agencies or individual investors. Some recent works in the literature study this changes in reporting financial news [32, 56, 57]. As Twitter usage in this context is definitively increasing; it is important to stress that, according to [50], stock microblogs exhibit three distinct characteristics about stock message boards: (1) Twitter's public timeline may capture the natural market conversation more accurately and reflect up to date developments; (2) Twitter reflects a more ticker-like live conversation, which allows micro-bloggers to be exposed to the most recent information of all stocks and does not require users to actively enter the forum for a particular stock; and (3) micro-bloggers have a strong incentive to publish valuable information to maintain reputation (increase mentions, the rate of retweets, and their followership), meanwhile financial bloggers can be indifferent to their reputation in the forum. Providing sensing, harvesting and analysing methods and tools of such information could be very useful for many stakeholders such as businesses and individuals making decisions to invest, stock market analysts and law enforcement agencies.

This paper analyses the permeability of Twittersphere, the total universe of Twitter users and their habits, to a relevant event for the financial market. Our hypothesis is that Twitter (although not a specific financial forum) is permeable to financial events and this permeability can be analysed by monitoring some specific features related to companies. Although several works in the literature have studied the correlation between Twitter and financial movements (to exploit its predictive power), our objective does not address regularities of the market in a longitudinal study that should take into consideration a long time series. On the contrary, we look for assessing the level of permeability of Twitter to financial events. For that, we study a remarkable event in January 2017: the announcement about the merge of Tesco PLC (hereinafter Tesco) and Booker Group PLC (hereinafter Booker) to create the UK's Leading Food Business on the 27th January 2017. Both companies, Tesco and Booker are listed in the main market of LSE (London Stock Exchange). The findings in the variation of Twitter volume and content show that the use of general purpose social media data is permeable to financial-specific events. The work presented is the first step in considering Twitter as a relevant feeder for taking decisions regarding the financial market and detecting irregularities in that market. An initial analysis of this experiment in [21] reported important insights when considering Twitter volume. This paper advances in the analysis of the Twitter permeability at that event by considering also the content of posts.

This paper is structured as follows. Section 2 introduces the related work the usage of digital information sources to assist decision-taking in the financial market, and Twitter in particular (Section 2.1). From this review of related works, Section 2.2 also highlights the contribution in this paper. In Section 3, we describe Twitter efforts to accommodate financial information in a general-purpose microblogging platform and we propose a model of the position of Twitter in the financial universe as well as the path flows of information. Also, we introduce the experiment and two specific research questions around our hypothesis. The experiment methodology (Section 4) takes advantage of the extraction characteristics of Twitter APIs to construct a specific dataset for the merger, which is described in Section 4.1 along with features related to the rapidness and synchronisation of Twitter with the LSE regarding the announcement of the financial event and the stock share prices (Section 4.2). Then, a series of investigations are conducted over the extracted dataset to address the research questions and the impact is analysed according to its disturbance in terms of Twitter volume (Section 5); in terms of hashtag dynamics and topic modelling (Section 6 and Section 7); in terms of sentiment towards the actions (Section 8); and in terms of geographical distribution of the posting (Section 9). Finally, Section 10 discusses our findings and

their limitations, meanwhile Section 11 summarises the main conclusions and introduces our ongoing work in the study of permeability of Twitter to financial events.

2 Related work

The modernisation and digitalisation of the financial market has been accompanied with the enormous increasing of information available for traders and investors, which can make extremely difficult for one person or a small group to conduct an analysis of a financial asset. In the application of computational intelligence to the financial market, mining textual content (financial news, financial reports , and even information in micro-blogs) is considered a relevant source of information for predicting future market behaviour [10]. To do so, researches have proposed the extraction of relevant features from the textual content. From the very beginning, sentiment was considered one of these relevant features to improve the accuracy of financial time series forecasting [54]. More recently, [27] proposes considering professional opinions in social media and financial news as supplementary sentiment sources, along with firm characteristics, so that, a methodology for joint effects of sentiment is proposed.

Cavalcante et al. [10] introduced a comprehensive review on the usage of computational intelligence in the financial market. The authors also suggest a systematic proposal for the building of intelligent trading systems. To sum up, most of the research in this field address the predictive power of social media problem and the differences appear in information sources, their extracted features, and also the predictive methods and the values to be predicted. [7] have shown that trading volumes of stocks traded in NASDAQ-100 are correlated with their query volumes (i.e., the number of users requests submitted to search engines on the Internet). Gunduz and Cataltepe [22] proposed a forecasting method which combines the analysis of news articles from Turkish finance websites, the extraction of feature vectors and stock prices to predict future market movements. The experiment used a naive Bayes algorithm to construct a prediction model which integrate the feature vectors and stock prices. Also, [35] used text mining of financial news-headlines to predict movements in the FOREX market; the extracted features include both semantics and sentiment in the content. Deep learning has been also applied to model both short-term and long-term influences of events on stock price movements in [17]. Finally, The combination of public news with the browsing activity of the users of Yahoo! Finance to forecast intra-day and daily price changes of a set of 100 highly capitalised US stocks was explored in [43]. The work showed that, when taken alone, sentiment analysis or browsing activity have very small or no predictive power and uncovered a "wisdom-of-the-crowd" effect that allows to exploit users' activity to identify and weigh properly the relevant and surprising news.

2.1 Twitter as a source for decision making

If the impact of information from online data sources into the financial market is broadly recognised by researchers and professionals, this is also the case for Twitter specifically. It is fair to say that it was Twitter that popularised the term hashtag as well as its # symbol to index keywords or topics so that people can easily follow those they are interested in. In 2012, Twitter unveiled a new clicking and tracking feature for stock symbols known as Cashtags. Cashtags are stock market symbols that can be included in tweets and, when preceded with a dollar sign (for example $VOD in regards to Vodafone), become clickable.

All the roles on the financial market (investors, specialised news agencies, etc.) are using Twitter to continually monitor the pulse of the market and make decisions. The literature pays particular attention to several ways in which the different market agents and participants may use Twitter analytics. Taking the plethora of works related to opinion mining on Twitter, one of the most researched areas is examining consumer behaviour for financial purposes. Although real-time decisions in the stock market are the most obvious manifestation of investment, long-term investments are more related with consumer analysis, so discovering the driving actors for sales and earnings is an active area of research [1]. Twitter is a valuable source of information even for the financial sector where contributors mainly fall into 5 categories [18]: (1) Journalists; (2) Companies and their representatives; (3) Government agencies; (4) Activist investors; and (5) citizen journalists (individuals). Also, the type of financial information is different, comprising not only of breaking news but rumours and speculations. According to [18], this new financial media comes with new challenges as the huge volume of available data, the high number of repetition of the same information and, notably, a continuum of quality of the tweets, demands mechanisms measure credibility. Also [12] suggests the importance of social media in the financial market, in particular Twitter, by analysing the popularity of Bloomberg tweets. Again, the authors agree on the fact that Twitter complements traditional news with speculations as well as off-the-cuff reporting, but also provides evidence that popularity within finance is not necessarily the same as popularity within other areas in Twitter, so that 'novelty' seems to be a very impacting feature of the popularity of financial tweets. Also, the importance of the very first source of the news itself, the negative disclosures in this case, has been analysed in [19]. Their experiment found evidence that negative financial news influences investors' willingness to invest when the news comes from the Investor Relations Twitter account, but not when it comes from the CEO's Twitter account. Taking apart the concerns about information veracity and credibility, freshness characteristics of Twitter also may have an important role in the field of High Frequency Trading (HFT), when traders make an investment position that is held only for very brief periods of time, even just seconds. At HFT, investors track social media to count for public behaviour and opinion to take their investor decisions. Thus, the relationship between Twitter sentiment and financial market instruments like volatility, trading volume, etc. and reports is investigated in [44] with promising results in DJIA and NASDAQ-100.

Hentschel and Alonso [23] reported an exploratory analysis of public tweets in English, extracted via Twitter Firehose, which contain at least one Cashtag from NASDAQ (National Association of Securities Dealers Automated Quotation) or NYSE (New York Stock Exchange). The analysis concludes that the use of Cashtag is higher in the technologic sector, which seems to be related to the technological profile of most of the Twitter users. In addition, the top 10 Twitter accounts according to the usage of Cashtags are companies or news agencies which in the majority of cases correspond to automatic or semi-automatic Twitter accounts. This analysis also highlighted the existence of relevant information behind the co-occurrence of Cashtags and the co-occurrence of Cashtags and Hashtags together.

Secondly, the analysis of the relationships between Twitter behaviour and stock share price is also a prime example of the increasing flow of information between financial and Twitter universes. For instance, [42] investigated a 15-month period of Twitter data including sentiment, concerning 30 stock companies registered on the Dow Jones Industrial Average (DJIA) index. This work gave some insights about sentiment and abnormal returns during the peaks of Twitter volume. Specifically, the authors show that not only is there a strong interaction between Twitter and the financial market in some moments identified as "known" relevant events (i.e. quarterly announcements), but similar results were observed in peaks not corresponding to any expected news about the stock market. Also, [29] used

Twitter to identify and predict stock co-movement according to firm-specific social media metrics and [49] studied tweets related to US market as indicator of some (potentially new) information in the stock market rather than on evaluating the problem of causality on stock prices.The results in [49] show that nearly a third of the tweets in the study are associated with abnormal price movements so that the authors suggest that Twitter is not a replacement for traditional sources in financial market. In fact, Twitter lacks the concrete trading recommendations that are common in other financial information sources.

Other research works have examined the possible connections between Twitter information and financial market performance, that is the predictive value of information gathered form social media to take decision about trading. Ruiz et al. [46] investigated the correlation between the activity on Twitter and financial time series, with the goal of verifying if published tweets can influence stock price movements or volume by using cross-correlation with time lags. Results showed that the trade volume of a stock was correlated with the number of connected components in the graph of that stock and with the number of tweets in the graph. However, the authors found that the price of a stock are weakly correlated with the analysed features. In [8] tick-by-tick transaction data was analysed for 20 Italian stocks on a period of approximately four months. Remarkably, [11] analysed lengthy dataset consisting of 1723 stocks for a period of more than two year, resulting in a dataset with approximately six million tweets which were tweeted by approximately 0.5 million unique users. In this study, authors found out that expert users impact financial market more than others and that sectors such as Technology and Customers show a better correlation than others with the financial movements.

Although most of these works are based on the Twitter data volume, there are also studies that apply sentiment analysis techniques in order to distinguish the polarity of content and its impact on the financial market. Bollen et al. [6] showed that public mood analysed through Twitter feeds is well correlated with Dow Jones Industrial Average (DJIA). Zhang [58] found out a high negative correlation between mood states like hope, fear and worry in tweets with the Dow Jones Average Index. Therefore, text based sentiment was considered useful to make trading decisions [2] or predict useful stock market variable, [13, 28, 36, 41]. Recently, [16] investigated the correlation of sentiments of public with stock increase and decreases using Pearson correlation coefficient for stock. Also [40] applied sentiment analysis and supervised machine learning to Twitter to analyze the correlation between stock market movements of a company and sentiments in tweets, finding out a strong correlation between rise/fall in stock prices of a company and public opinions or emotions about that company expressed on twitter. As it is widely accepted, one of the main issues in natural language processing in Twitter is the short length of the posts that might limit unsupervised learning. However, [37] proposed an approach for creating stock market lexicons from the specialised stock market microblog StockTwits. Then, unsupervised machine learning is applied to produce Twitter investor sentiment indicators. The results shows a moderate correlation with two traditional survey indicators: Investors Intelligence (II) and American Association of Individual Investors (AAII). Same authors proposed in [38] a methodology based on Kalman Filer to forecast stock market variables: returns, volatility and trading volume of diverse indices and portfolios.

2.2 Contribution

Consequently, there is a general consensus of the good sensing and novelty characteristics of Twitter as a source of information for the complex financial market. However, the provenance of contributions, their different levels credibility and quality and even the purpose or

intention behind them makes Twitter not reliable enough as the single source for decision taking. This paper reports a Twitter study framed by a collaboration project among the University of Vigo and the Manchester Metropolitan University to deploy an architecture for real-time monitoring of irregularities in the stock market. That architecture will apply data mining and fusion technologies from a pool of social media feeds related to the stock markets. In order to design the architecture, the permeability of the different feeders should be analysed, that means, to what extent a specific financial information feeder is permeable to fraudulent and common irregularities in the financial market. The aim of this paper is to analyse that permeability for the case of the microblogging platform Twitter. The Intelligent System Group at the Manchester Metropolitan University has worked in the detection of irregularities form Financial Discussion Forums [39]. Meanwhile, the Information & Computing Lab of the University of Vigo has a relevant know-how in applying Twitter analytics to a variety of real life problems [45, 48].

Meanwhile the vast majority of the studies focus on the predictive power of social media regarding the movements of the financial market, the work in this paper focus on anomaly detection. Also [4] addresses the problem of analysis specifically how investors on Twitter behave around the meeting dates of FOMC (Federal Open Market Committee) USA Federal Reserve System. They found that the polarity score of tweets that are published at least 24 hours before the FOMC meeting are relevant to predict returns. So, this paper looks at social media's reaction to a significant and recurring macroeconomic event, instead of analysing events for individual equities. In our work, we are interested in research the reaction to significant financial events that barely can be analysed with longitudinal studies, which explore the regularities of the market along a long period.

3 Research questions and the experiment

But, Twitter is neither the only nor the biggest source of information about financial markets and, in this respect, Figure 1 shows a reference model for the Twitter position regarding this universe of information and its relationships within the stock market. At any moment (time) and from anywhere (location), a variety of contributors in the financial sector (verified and unverified Twitter accounts) may post (tweet or retweet) pieces of information referring to a company by using either (1) the unique symbol $cashtag, (2) #hashtags related with the company for a specific purpose or related to a specific event and /or (3) simply the company name in its different forms (official, abbreviation, colloquial, etc.). At the same time that main agents in the financial market post into Twitter, the Twitter content maintains references to the financial universe outside via mentions, URLs, etc. Given apart the flows which characterise the permeability of the layer in between Twitter and the financial market, the information naturally spreads throughout the Twittersphere by the common re-tweeting and following mechanism in the platform. Even if the piece of news (post) comes from a human spread rumour or some cyber-attack which ends up being false information, [26] sustains about "the power of networked social and financial systems to connect autonomously and to produce a present without human oversight or governance". In our opinion, further actions and research are needed to achieve a safer financial ecosystem for traders and investors.

In this section, we describe the experimental methodology which will be used to answer our main hypotheses: Twitter (although not a specific financial forum) is permeable to financial events and this permeability can be analysed by monitoring some specific features related to companies. To address this high-level research question, we selected a random but relevant even, according to its impact on the financials spheres but also on the general public,

at the beginning of year 2017. The most important feature for the selection of the event is the latter, that is, an event which being intrinsically financial, is of common interest for the general audience so that permeability can be measured beyond the spheres of financial experts in Twitter.

The event under study, is the merge of Tesco and Booker on the 27th January 2017, announced by the RNS (Regulatory News Service) of LSE (London Stock Exchange) at 7:00 a.m. GMT (Greenwich Mean Time). This action is modelled according to the reference model in Fig. 1, that is, we model Tesco on Twitter with the conceptual triplet (cashtag, hashtags, keyword), that is ($TSCO, #[Tesco], "Tesco"). This model represents the financial perspective of Tesco on Twitter ($TSCO), specific comments about some Tesco issue on Twitter (#[Tesco]: #hashtag in some tweet with the keyword "Tesco"), and general references to Tesco on Twitter ("Tesco"). As mentioned, our objective is a monitoring platform that continuously harvests signs scattered all over social media platforms to identify irregularities, such a platform will be supported by a data fusion model which extract relevant features and weight the contributions from different online information sources. Although in such an approach not only Tesco but Booker would be continuously monitored, the experiment faces the isolated behaviour of Tesco (one of the parts in the action) due to significance reasons, as the part with more presence in social media and due to design reasons, as the company being the focus of real-time irregularity monitoring.

Regarding permeability to financial events, we may hypothesise that the permeability and the impact is not alike in the three perspectives which constitute the triplet. The cashtag is invariably linked to financial news of a company but hashtags have a completely different dynamic. The hashtags related to a specific company will emerge and disappear dynamically according to the company decisions, marketing campaigns, consumer behaviour, etc. Tesco being a well-known company in the UK, the impact of financial events in hashtags might be limited and just visible in case of general-public financial events. Presumably, financial events should have a bigger impact on cashtag tweets (according to volume and context) than on hashtags (just altering their dynamics) or on topics of Tweets content. Nevertheless, this presumably different behaviour should be inspected.

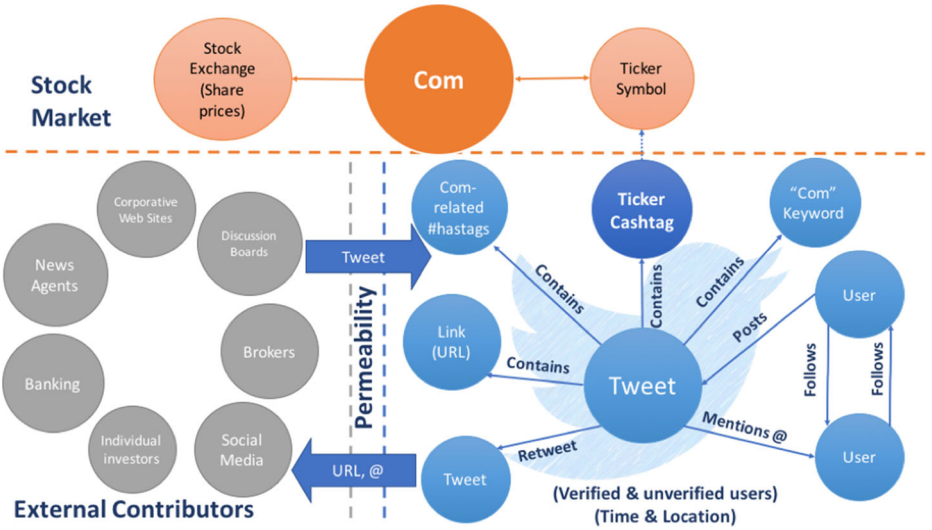


Fig. 1 Twitter position in the universe on financial information

Given our hypothesis: ***"Twitter (although not a specific financial forum) is permeable to financial events and this permeability can be analysed by monitoring (1) the name of companies as a keyword ("Tesco" in this case), (2) the Cashtag of the company ($TSCO) and (3) the hashtags related to that company."***

The experiment will attempt to answer the following research questions (RQ) related with the permeability of Twitter to financial events:

- RQ1: Is there any difference in permeability, in terms of volume and dynamics, among $cashtag-content (specifically $ marked financial tweets) and #hashtag-content (general tweets with #hashtag marks) and financial content (general tweets with financial terms)
- RQ2: Is that permeability (in terms of volume and dynamics) accompanied by additional variations in some other features in the content (location, sentiment)?

On the basis of the proposed RQs, and even though we are reporting a single case study, we would like to highlight that our triplet approach can be applied to the general monitoring of financial event on Twitter disregarding the market and the sector if we consider announcement services and stock prices in other regulated financial markets.

4 Twitter data mining & the dataset

There are three different ways to obtain Twitter data: Search API, Streaming API and Firehose. The Twitter Search API provides the endpoints to recover tweets that were published in the previous two weeks, with the possibility of filtering according to several criteria. On the other hand, Twitter Streaming API returns 1% of the tweets that match some search parameters in real time. Finally, Twitter Firehose provide access to the 100% of the tweets, but it is not a free-access API. Twitter APIs are constructed around four main "objects": *Tweets*, *Users*, *Entities* (hashtags, URLs, mentions and media in a tweet) and *Places*. From theses object and taking time as the Twitter's backbone, we consider a Twitter model which contemplates 3 orthogonal perspectives: The content, the social structure and their spatio-temporal context (see Fig. 2).

The anatomy of these objects is described in the Twitter Developer documentation. With regard to this work, as the experiment does not focus on the dynamic spreading of information on Twitter (via retweeting, mention or like mechanisms), we select the following features (existing in both APIs under different field names) for the analysis, all of them accessible from a Tweet object:

- Content perspective: the status update (`Tweet:text`) and the entities in the tweet (`Tweet:entities`), specifically `hashtags` (including cashtags) and `urls`.
- Context perspective: the post time of the status update (`Tweet:created_at`) and, if available, also the place by feature `Tweet:coordinates` and feature `Tweet:place:bounding_box`).
- Social Perspective: User (`Tweet:user`, specifically the field verified).

There are some differences between the Searching API and the Streaming API illustrated in Fig. 2. One difference being the time direction - the most relevant one to our experiment. The Search API goes back in time, whilst the streaming API goes forward. Moreover, there are other differences related to mainly the format and the rate limit rules. As it is shown in Fig. 2, the search and Streaming API does not return data in exactly the same format but the differences in format are irrelevant for data analysis where preprocessing can define a

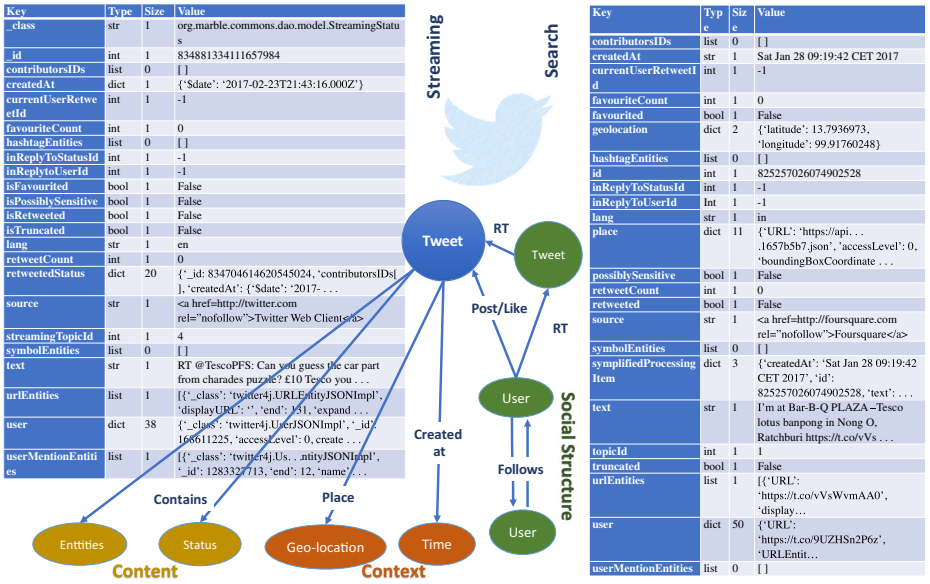


Fig. 2 Twitter APIs and a model of the Twittersphere

uniform intermediary format. Regarding their extracting capacity, forums contain plenty of discussion about this issue which has not ever made enough clear from Twitter officially.

According to the proposed model (Fig. 1) the experiment will analyse the permeability of Twitter to financial events by the inspection of the triplet ($TSCO, #[tesco], 'tesco'), defined as:

- $TSCO, the set of tweets where the ticker symbol $TSCO is an entity;
- #[tesco], the set of tweets with at least one hashtag and containing the keyword 'tesco'; and 'tesco',
- the set of tweets containing the keyword 'tesco', no matter whether they contain a hashtag or not.

4.1 The dataset

The extraction strategy firstly used the Twitter Search API to recover the information backwards before the announcement on 27th of January 7:00 a.m. GMT and the streaming API was used to recover information forwards until 27th February (one month later). This data streaming collection, just after the announcement, was used to visualise the impact of the announcement and the time the behaviour of tweets concerning Tesco resumes a regular pattern again. The results of the combination of the search and streaming results is shown in Fig. 3.

Once the behaviour becomes regular, we used the Search API again to obtain a regular dataset to compare with the Search API results recovered just after the merge action was announced. Clearly, the Twitter Search API is not appropriate for continuous analytical monitoring and as a data source to allow real-time decision making. It is not intended and does not fully support the repeated constant searches that would be required to deliver 100% coverage. However, the experiment in this paper is limited to one individual company, 2 keywords and timelines in the scale of weeks. Therefore, the Search API provides a better

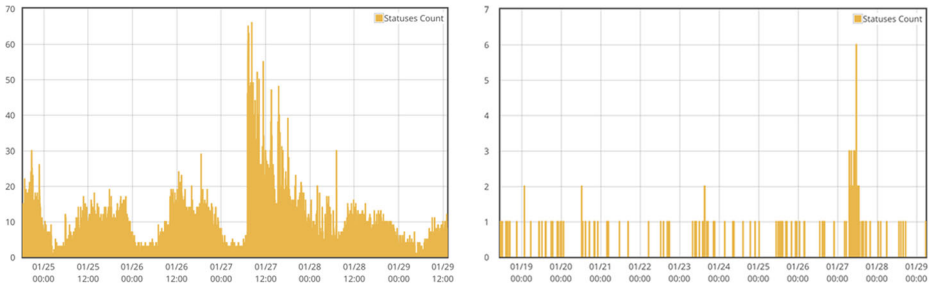


Fig. 3 Total Twitter volume for the keyword 'tesco' (left) and for the cashtag $TSCO (right) by merging (without duplicates) data returned from queries to Search API (backwards) and Steaming API (backwards)

coverage than the Streaming API (1% according to the Twitter official information), if the superior filtering characteristics of the Search API are used. Nevertheless, as the Search API has a limit on the number of tweets recovered, to get the whole data during the period under study (see Table 1), we repeatedly asked Twitter for the most recent results backwards by windowing the searches according to the publication date and merging results according to the post Id. In this way, we guarantee a fair comparison according to the volume of data since, in any manner, we should compare the Search API with Streaming API results. According to that, and to give response to the research questions, we use the Search API queries to cover the time periods defined in Table 1. Mainly, the 4 periods in this table capture the Twitter volume retrieved by the Search API during the very same days of the week, Wednesday to Sunday, around the announcement on Friday 27th of the merge (pre-announcement and post-announcement periods) and when behaviour becomes regular at the corresponding week days. Two types of Search API queries were considered, (1) query with the term 'tesco' for the Tweet:entities and the Tweet:text entities, to capture post related with Tesco and post also containing a hashtag related with Tesco (a total of 70,793 tweets) and (2) query with the term '$TSCO' for Tweet:entities (a total of 151 tweets). It is fair to mention that 'tesco' volume is several orders of magnitude higher than '$TSCO' given the high visibility of the company Tesco in social media as a vehicle for marketing and consumer engagement.

Table 1 Time Periods (GEM Time) in the experiment (extracted with Twitter Search API)

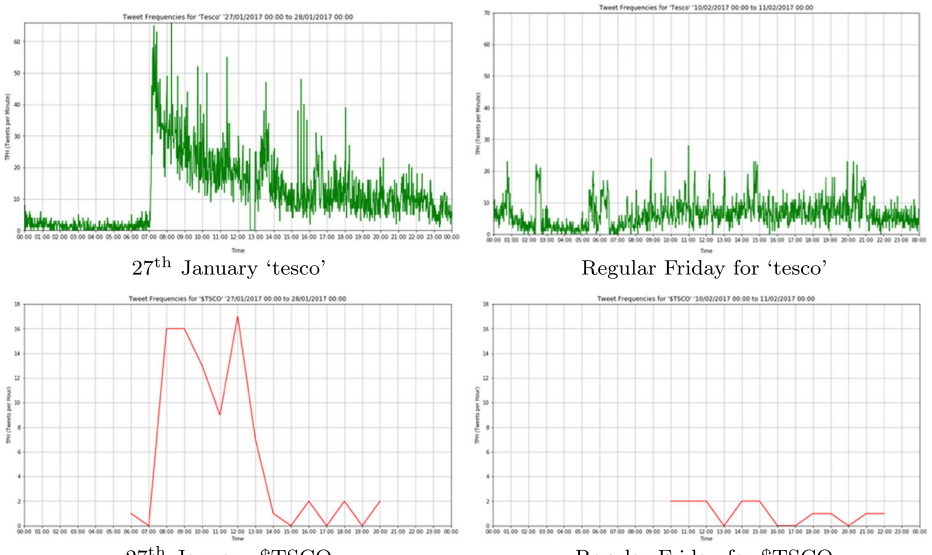
Name/Period	'tesco'		$TSCO	
	incl. #[tesco]			
	Total	Per/hour	Total	Per/hour
Pre-announcement				
25th Jan 00:00- 27th Jan 07:00	11,817	214.85	12	0.218
Post-announcement				
27th Jan 07:00 - 29th Jan 23:59	25,547	393.03	91	1.4
Regular 2-weeks-after				
8th Feb 00:00-10th Feb 07:00	13,417	243.94	26	0.473
Regular 2-weeks-after				
10th Feb 07:00 - 12th Feb 23:59	20,012	307.88	22	0.338

'tesco' dataset: Fri Jan 27 07:03:59 2017 $TSCO dataset Fri Jan 27 07:03:46 2017

Fig. 4 First Tweet referring to the action in the 'tesco' and $TSCO dataset

4.2 Rapidness & synchronization in the dataset

Although our analysis focuses on the permeability of Twitter to financial events, our objective, and part of our future work, is the use of Twitter as a sensor of irregularities in the stock market so rapidness and synchronisation of Twitter as a channel to the stock market: rapidness in its response to the RNSs of LSE (London Stock Exchange) and



27^{th} January 'tesco' Regular Friday for 'tesco'

27^{th} January $TSCO Regular Friday for $TSCO

Fig. 5 Time series at a minute scale on the 27th January in comparison with a regular Friday

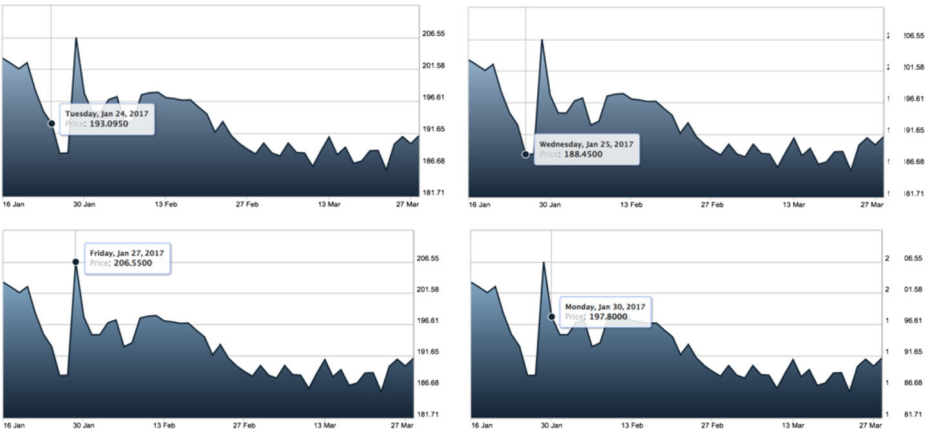


Fig. 6 Main observational points in the Evolution of the Tesco PLC share price (16th January to 27th March 2017)

synchronisation with the share prices also in LSE are relevant. Regarding the rapidness, the experiment definitively shows the good characteristics of Twitter. The first tweet referring to the RNS was at 7:03 a.m. on 27th, just 3 minutes before the RNS announcement about the Tesco and Booker merge (Fig. 4). Beyond the very first tweet, it is remarkable the rapidness of the peak response to the announcement in both datasets, so that the 27th Twitter time series ('tesco' and $TSCO) can be considered abnormal time series when a regular Friday is taken as a reference. We highlight that the peak starts from 7:00 to 8:00 both in the #TSCO and $TSCO dataset (see Fig. 5).

Regarding the synchronisation with the share prices at LSE (Fig. 6 and Table 2), it is fair to mention that although the share prices were abnormally low the day before the announcement, we haven't found any reference to the financial vocabulary considered for the action during this period. Moreover, we cannot extract a sound finding given to the concatenation of two events with financial impact in Tesco PLC: the legal actions against Tesco PLC overstatement (see Section 7) and the merge of Tesco and Booker. In terms of rapidness and synchronisation, we appreciate that the Twitter permeable layer exhibits rapidness in response to the financial market but we barely appreciate a clear synchronisation with the stock share prices.

Table 2 Values of share prices and Twitter volume

Period	Price closing time	'tesco'	$TSCO	Total
24th 16:30 to 25th 16:30	188.45	7,697	6	7,703
25th 16:30 to 26th 16:30	188.5	8,356	4	8,360
26th 16:30 to 27th 16:30	206.55	15,295	38	15,333
28th 16:30 to 29th 16:30	—	10,077	8	10,085
29th 16:30 to 30th 16:30	—	4,539	3	4,542

5 Impact on twitter volume

In this section, we detail the impact of the event by analysing the variation in the number of tweets (volume) so that a quantitative measure of Twitter permeability to a financial event can be observed. During this part of the analysis some irregularities were discovered which related to an inconsistency in the named scheme of tickers in Twitter. In particular, to our knowledge, Twitter has not promoted the specific distinction among financial markets so that the uniqueness of ticker symbols inside a market disappear in the Twittersphere. That is the case of $TSCO cashtag which corresponds to Tesco PLC in the LSE and to *Tractor Supply Company* in the NASDAQ, the second stock exchange in USA. So, the returned results to a $TSCO query included tweets related to Tesco PCL and also to Tractor Supply Company. If the cashtags is the entity to aggregate information around a specific company and, consequently, to allow the spreading of such information, some kind of market prefix should be used, especially in the times when companies are increasingly global.

Figure 7 shows the temporal series in tweets per hour (TPH) scale. Although it is quite obvious that the number of TPH in the 'tesco' dataset is up several orders of magnitude higher than those of the $TSCO dataset, the peak behaviour is more acute in the $TSCO

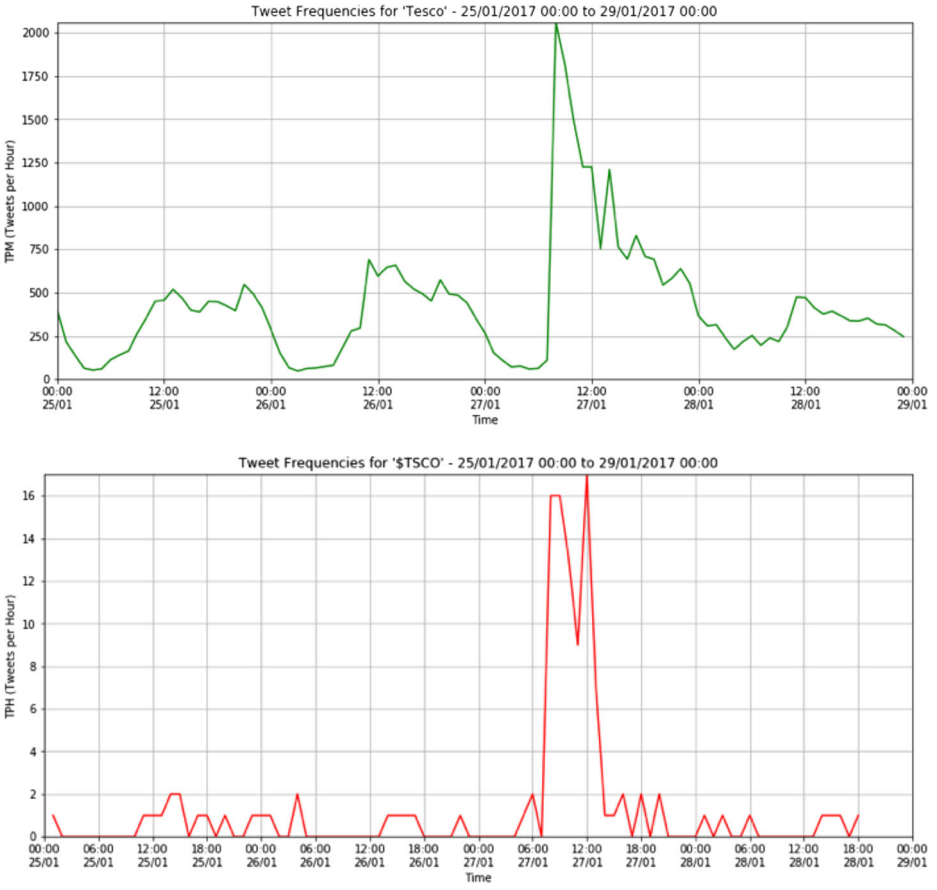
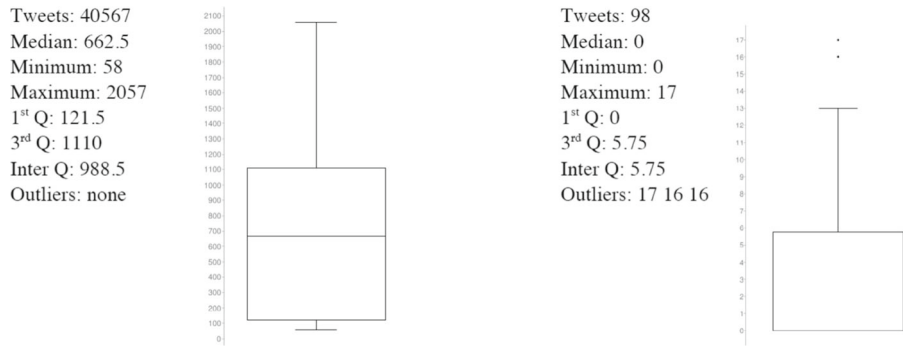


Fig. 7 Time series of the 'tesco' and $TSCO dataset from 25th January to 29th January

Table 3 Peak behaviour on the 27th January for 'tesco' and $TSCO



one. As it is shown in Table 3, considers the hourly volume of the 'tesco' dataset on the 27th January. There are no outliers during the day, with the peak value of 2,057 tweets occurring in the sample from 8:00 to 9:00. Nevertheless, there are 3 outliers in the $TSCO dataset: samples 8:00-9:00, 9:00-10:00, corresponding to the time just after the announcement and 12:00-13:00, being consistent with previous studies about social timing in [34], which showed peak activity during lunch time in different cities around the world.

Apart from the peak comparison, we inspect the disturbance on other dataset features before and after the event, also comparing these dates with the regular behaviour 2 weeks later. In Table 4, we compare the behaviour of the main features of 'tesco' data (in green) and $TSCO' data (in blue).

Firstly, we can observe the increase of Tweets per hour during the post-announcement, compared to the regular period, is more acute in the $TSCO data (1.40 vs 0.34) than in the 'tesco' data (393,03 vs 307.86).Secondly, the percentage of tweets which contain a URL are significantly higher in the $TSCO data (from 75% to 82.61% in the different periods) with

Table 4 Variability of features in 'tesco' and $TSCO data before and after the announcement in terms of regular behaviour

Periods	**Pre-Announcement (25th-27th Jan 06:55) TUESDAY-WEDNESDAY-THURSDAY**				**Post-Announcement (27th 07:00- 29 Jan) FRIDAY- SATURDAY-SUNDAY**			
Counting & percentages	**"tesco"**	**%**	**$TSCO**	**%**	**"tesco"**	**%**	**$TSCO**	**%**
Tweets	**17,154**		**8**		**25,547**		**91**	
Tweets per hour	**311.89**		**0.15**		**393.03**		**1.40**	
Tweets from verified users	2,560	14.92%	0	0.00%	2,696	10.55%	2	2.20%
Tweets with URL	5,383	31.38%	6	75.00%	12,367	48.41%	64	70.33%
Tweets being RT	7,522	43.85%	0	0.00%	8,070	31.59%	18	19.78%
Different users	12,141	43.85%	7	0.00%	15,757	31.59%	47	19.78%
Different erified users	155	1.28%	0	0.00%	336	1.32%	2	4.26%
Periods	**2 Weeks after (7th - 9th Feb 06:55) TUESDAY-WEDNESDAY-THURSDAY**				**2 weeks after (10th 07:00 - 12th Feb) FRIDAY-SATURDAY-SUNDAY**			
Counting & percentages	**"tesco"**	**%**	**$TSCO**	**%**	**"tesco"**	**%**	**$TSCO**	**%**
Total Tweets	**16,878**		**23**		**20,011**	**100.00%**	**22**	
Tweets per hour	**306.87**		**0.42**		**307.86**		**0.34**	
Tweets from verified users	2,364	14.01%	0	0.00%	2,650	13.24%	0	0.00%
Tweets with URL	4,980	29.51%	19	82.61%	6,971	34.84%	19	86.36%
Tweets being RT	6,530	38.69%	3	13.04%	5,676	28.36%	5	22.73%
Different users	10,749	38.69%	10	13.04%	11,374	28.36%	13	59.09%
Different verified users	164	0.97%	0	0.00%	150	0.75%	0	0.00%

respect to the 'tesco' one (from 29.51% to 48.41%), which is a result of the professional and financial orientation of the $TSCO data as a channel to spread facts and news rather than opinions and sentiments. Finally, the retweeting activity is higher in the announcement periods (pre- and post- with 43.85% and 31.59%) compared to the regular periods (38.69% and 28.36%) for the 'tesco' data. The increase of retweeting is, by nature, linked to the need or desire of spreading a piece of content but, the reason behind may be different as, in fact, it is in our case study: retweeting the 'tesco' keyword is mainly related with a Tesco campaign involving retweeting (see section VI for the details) meanwhile retweeting in $TSCO data is mainly linked to spreading the information about the merge (post-announcement) and about other financial news related with Tesco PLC. The information about this retweeting activity is expanded upon in the following sections of this paper. Finally, we observe the invariability on the number of verified users either along all the periods and along the 'tesco' and $TSCO data.

Both the peak behaviour (Table 3) and the feature comparison (Table 4) provide an answer to **RQ1**, that is, the event impact on $cashtag-content is greater than on #hashtag-content or on general content.

6 Impact on hashtags dynamics

Given the volume of Tweets, the spreading of the event throughput the Twittersphere should presumably have two effects: the change on the hashtag dynamics and the change on the main topics in tweets. More than this effect, what we would like to measure is to what extent this disturbance can be perceived differently in 'tesco' and $TSCO data. Presumably, and contrarily to the volume analysis, the content impact is expected to be more marked or noticeable in the 'tesco' data. So that, financial topics, which are the core of cash-tagged tweets, would conquer the whole Tesco content as a result of the good permeability to this financial event in Twitter, should be studied.

Focusing on the analysis on hashtags as a measure of ephemeral or nor so ephemeral interest, Table 5 shows the disturbance of the event in hashtag dynamics in the 'tesco' data, which is considered data related with the company without any specific financial bias. In terms of the 5 most frequent hashtags during the periods considered in the experiment, we clearly perceive that, only during the period Post-Announcement, financial hashtags emerge: #teschoshareprice and #booker. We consider #booker a hashtag related with the action given its co-occurrence with the keyword 'tesco', the reverse may be not true. Also, we discover a persistent hashtag during all the periods: #essothursdays, related with the equally named promotion campaign which allows Twitter and Facebook users to obtain 100 Tesco voucher provided that they follow the Twitter account @GB_Esso and retweet the competition tweet.

Regarding hashtag disturbance at a more general level, the bottom part of Table 5 shows the percentage of occurrence of the main hashtag of the company (#tesco), hashtags related with main business of the company (#-main business), and hashtags related with merge and financial aspects (#-action and financial). Also, Figure 8 shows hashtags in terms of semantic families. The left chart shows the main topics reported in the 'tesco' data and the right chart distinguishes the three main categories within these topics: "Tesco" in grey, "Main Business" in yellow and "Action and Financial news" in green). It is clear that, from the point of view of the hashtags dynamics, the merge reporting achieves a high position in the 'tesco' dataset after the announcement. During the pre-announcement period, hashtags (apart for the generic one *#tesco* in grey) are related to marketing campaigns (yellow in the

Table 5 Disturbance on Hashtag Dynamics for the 'tesco' dataset

Pre-announcement		**Post-announcement**		**Regular (2 weeks after)**	
3171		1407		1917	
Hashtag	Occurrence %	Hashtag	Occurrence %	Hashtag	Occurrence %
#essothursdays	78.27 %	#tescoshareprice	40.80 %	#essothursdays	49.67 %
#tesco	12.44 %	#essothursdays	23.81 %	#tesco	16.15 %
#sidedish	3.26 %	#bbcbreaking	15.71 %	#valentinesday	10.95 %
#deal	3.19 %	#booker	11.16 %	#win	12.09 %
#bagsofhelp	2.85 %	#nieuwstwitter	8.53 %	#freebiefriday	11.14 %
Hashtag Topic	Occurrence %		Occurrence %		Occurrence %
#tesco	12.44 %		0 %		16.15 %
#-main business	87.57 %		23.92 %		83.85 %
#-action and financial	0 %		76.08 %		0 %

aggregated chart on the right). It is only during the post-announcement period that financial-related hashtags emerge: *#tescoshareprice* being 40.80% and #booker being the 11.16% of the top hashtags. Also, it is remarkable that the emergence of the hashtags related to the breaking news *(#bbcbreaking* and #nieuwstwitter) and that even #tesco disappears in the top-5 hashtags during the post-announcement period (Table 5). All of financial-related and breaking news hashtags completely disappear in the period we consider as regular in terms of Tesco experiment (see also Table 5).

Lastly, regarding the analysis of hashtags in the $TSCO dataset, it is not strange than during the regular period none of the tweets in the dataset contains a hashtag, so no extra specific information to follow and spread information is provided, apart from the pseudo-hashtag $TSCO. However, during the post-announcement period 15,38% of the tweets in the $TSCO dataset contains at least one hashtag. Nevertheless, the number of tweets during this period is 91 of which 16 include at least one hashtag. Although we prefer not to analyse the hashtag dynamics with this small number, we can appreciate a disturbance with regards to the regular period. This result talks about the malfunction of #hashtags in the $TSCO dataset. If #hashtags are the Twitter tool to aggregate semantically-related content and spread it, this feature is almost absent in the $TSCO dataset. We cannot perceive a hashtag disturbance related to the action, because in the financial island of Twitter the $cashtag

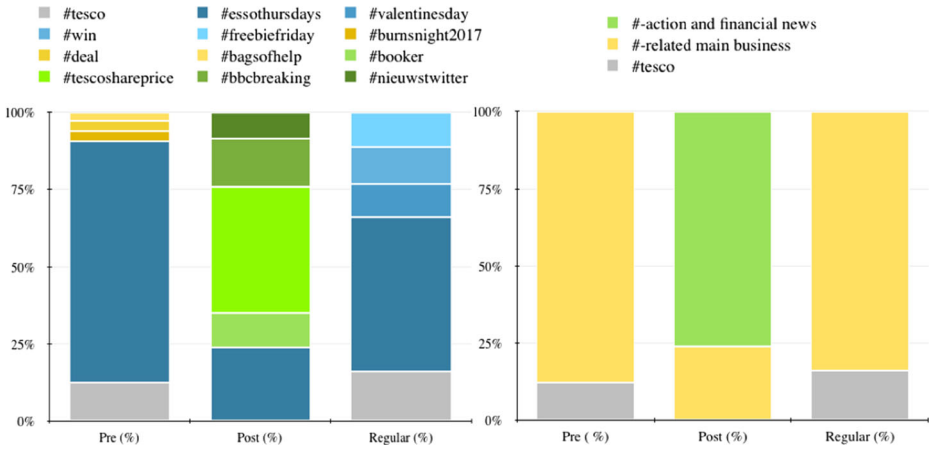


Fig. 8 Disturbance on Hashtag Dynamics for the 'tesco' dataset (left) and aggregation of the hashtags in semantic families

is the main Twitter tool to aggregate and spread content. What we can only appreciate with respect to the regular behaviour is the increase on the number of hashtags which accompany the pseudo-cashtag $TSCO.

According to these results we can establish the fulfilment of RQ1, in that Twitter permeability to the merge can be perceived not only on $cashtag-content, which has a clear financial orientation, but also in the #hashtag dynamics.

7 Impact on the vocabulary & topics

When analysing the impact of the action on the Twitter content, the complete response to **RQ1** (can a financial load be also perceived in the #hashtag dynamics as in the general content?) should be studied also from the point of view of the main topics people post about Tesco. For that, we should extract these topics from the tweets content in 'tesco' data and compare them with the ones in the specifically financial $TSCO data. Unfortunately, the number of tweets in the $TSCO dataset is relatively low. According to this, we decided to automatically split the 'tesco' and $TSCO data according to an automatic financial annotation by using a simple vocabulary related with general financial terms and specific terms related with the action.

The first RNS related to the event was published 27th January 2017 as RNS number: 2907V under the title "TESCO & BOOKER ANNOUNCE MERGER". The body of the announcement contains the following preface "The boards of Tesco PLC ("Tesco"), the UK's leading food retailer, and Booker Group plc ("Booker"), the UK's leading food wholesaler, are pleased to announce that they have reached an agreement on the terms of a recommended share and cash merger (the "Merger") to create the UK's leading food business." To measure the financial burden of the topics in tweets content and the disturbance of this burden before and after the announcement, we apply a priori knowledge by selecting a simple vocabulary by inspecting the common terms in the content of the $TSCO dataset and the content of the RNS. Two categories of terms were created:

- Generic financial terms: Board of Directors, Executive Manager, CEO, Chairman, Director, Share, shareholder, revenue, return, capital, investment, dividend, Mix & Match.
- Specific action terms: Merge, take over, announcement, RNS, Merger, Tesco, Booker, Combined Group, Stewart Gilliland (extracted for the RNS announcement and the main actors in the action. The announcement is 85 pages long).

Regarding the representativeness of our vocabulary, the frequency of terms in the 'tesco' data was 7.11% and is much lower than the frequency in the $TSCO data which is 47.68%. We remark that these percentages referred to the total corpus (content of all the tweets in the dataset), not the total number of tweets in the datasets. Regarding the number of tweets containing at least one term in the vocabulary, we observe the following results in 'tesco' dataset: 2.07% (245 of 11817) before the announcement and 9.52% (2437 of 25547) after the announcement. This difference is obviously not so big in the $TSCO dataset: 25.00% (3 of 12) before the announcement and 27.47% (25 of 91) after the announcement (the increasing can be explained because of the inclusion of terms specifically related with the action). These differences in percentages is what these datasets should presumably exhibit according to the clear financial orientation of $TSCO as a pseudo-hashtag for the financial facts related with Tesco PLC. So that, our simple vocabulary allows us to measure the financial load of a set of tweets.

According to the vocabulary, we split the 'tesco' dataset (pre- and post- announcement) into a set of tweets containing at least one term in the vocabulary, hereinafter referred as *'tesco' financially-oriented subset*; and the set of tweets not containing any term in our vocabulary, hereinafter referred as *'tesco' general subset*. Then, for uncovering the main topics, we adopted a simple bag-of-words model [55]. Under this approach tweet words in the status content are only considered according to their relative frequency and not according to their order within the document. In the bag-of-word result, we consider the proper cleaning precautions (i) by lowercasing and removing stop words; and (2) by specifically considering Twitter jargon (i.e. HT, QOTP). Finally, terms are contextualised according to the tweet semantics and categorised into topics. As a result, we obtain the contextualised & categorised occurrence of terms in Table 6, and the evolution of the size of the considered categories on the 'tesco' corpus in Fig. 9.

More specifically, Table 6 contains the contextualisation of the 20-most frequent terms inside the complete content of the tweet according to an n-gram strategy and manual annotation. We can obviate the term 'tesco' in its different forms (tesco, tesco's in grey) from the analysis, since the dataset has been obtained from different 'tesco' search queries to Twitter. Regarding the general annotated subset of the 'tesco' dataset, before 27th 7:00 a.m. all most frequent terms are either stop words (including Twitter jargon) and terms related with the main Tesco PLC business. Even though, we can distinguish a set of terms related with a specific Tesco campaign (the one previously mentioned in the appearance of #essothursdays in blue) and also a topic 'January blues' which can be considered seasonal in the UK, even in the context of a dataset specific for the company Tesco PLC. Just after the announcement, the scene totally changed and neither the campaign nor the seasonal topic appear but they are replaced by terms related to the Tesco PLC merge with Booker. Therefore, topic discovering exhibits the same kind of disturbance than hashtags, that is, the impact of the merge can be perceived in the *'tesco' general subset* after the announcement. After this analysis, we can establish the fulfilment of the complete **RQ1** : Given that $cashtag-content has a clear financial orientation, this financial burden can also be perceived in the #hashtag dynamics and in the general content. So that, Twitter is highly permeable to the financial event considered in this experiment.

Regarding the *financially-oriented subset* of 'tesco', the most frequent terms are obviously related with Tesco PLC financial perspective before and after the announcement but,

Table 6 Term occurrences in subsets of the 'tesco' dataset (General and Financially-oriented)

Pre-announcement 'tesco' dataset						Post-announcement 'tesco' dataset					
'tesco' general subset			*'tesco' financially-oriented subset*			*'tesco' general subset*			*'tesco' financially-oriented subset*		
Topic	Term	Occurrences	Topic	Term	Occurrences	Topic	Term	Occurrences	Topic	Term	Occurrences
tesco	tesco	9,527	tesco	tesco	165	tesco	tesco	16,123	tesco	tesco	2,089
Main Business	save	1,018	Main Business	return	63		tesco's	864		wholesaler	583
	points	751		stock	140	Main Business	food	3,010		booker	1,606
	t&cs	2,255		store	55		deal	3,543		billion	261
	voucher	2,329	Over-statement	action	92		news	872		3.7	238
100 voucher GB-esso	100	2,322		scandal	72		wholesaler	3,147	£3.7 billion deal	deal	687
	win	2,365		shares	59		3.7bn	2,675		announces	239
	gb	2,288		bt	50		booker	7,856		merger	1,307
	essothursdays	2,282		joins	50		group	2,567		merge	467
	chance	2,273		vw	50	£3.7 billion deal	uk's	1,332		agreement	271
	ffw	2,263		crosshairs	50		londis	1,324		group	278
January blues	blues	2,259		shareholder	50		budgens	1,313		3.7bn	545
	january	2,297		specialists	46		merger	1,313		uk's	106
Stop words	please	741		report	45		biggest	1,263	Financial news	news	189
	back	990		accounting	29		owner	944		breaking	112
	i'm	892		legal	96		billion	920		shares	294
	hi	819	Undetermined	new	48	Undetermined	buy	3,653	Undetermined	food	410
	life	1,057	Stop words	hi	40	Stop words	i m	1,212		reached	175
	gone	2,339		please	30		back	879		stock	158
Twitter	ht	2,237		i'm	27		this	869		buy	172

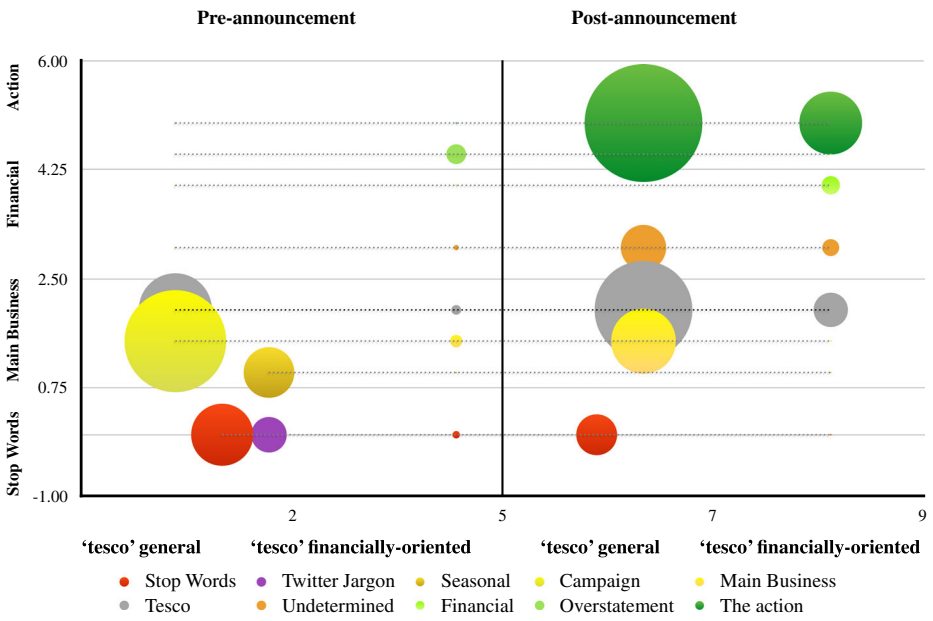


Fig. 9 Evolution of Main topics in the 'tesco' corpus

once again, the scene changes totally from one day to another. With the financial lens the filtering vocabulary provides, a topic emerges referring to another relevant financial event related with Tesco PLC. On Jan 24th, a Tesco spokesman announced that the firm was facing a claim for damages about profit overstatement. This topic is, therefore, relevant enough to emerge in the financial subset but is disguised by the contents related with the main business in the general subset. What we have is a financial event, in the same company, which does not surpass the imaginary threshold of permeability to the general Twittersphere. This can be clearly perceived in the Fig. 9 where bubbles are coloured according to the main topics (green scale for financial and yellow scale for main business, grey for Tesco in the middle of these two worlds) and sized according to the number of term occurrences. A big green bubble appears in the general post-announcement subset but, in the financially-oriented subset, the green bubble related to the overstatement is replaced by the green bubble related to the merge.

8 Impact on polarity

Unlike topic modelling, sentiment analysis is about applying natural language processing (NLP) to mine the subjective impression beyond the actual facts and to measure the individual sentiment or reactions toward certain products, people or ideas by revealing the contextual polarity of the information. The python library *TextBlob*, [31] was used to discern how positive or negative the sentiment in the tweet is. It is worth stating at the outset that we do the analysis with the same polarity classifier in two quite different datasets, the more general 'tesco' data, which contains plenty of sentiments, and the financial $TSCO dataset under the lens of the cashtag, more oriented to facts or as much opinions than sentiment behind human beings. Also, we used a polarity classifier previously trained on general

Tweets and not specifically trained for financial content and its peculiar jargon. With all these precautions in mind, and taking into account that we are facing a single event, we provide an interpretation of the sentiment analysis (Table 7) supported by the knowledge acquired throughout the previous sections.

Regarding the 'tesco' dataset, the analysis during the regular period in the experiment exhibits a 33.19% positive, 44.54% neutral and 22.27% negative sentiment load. Contrarily, just before the announcement the same dataset increases its positive sentiment to 55.2% which seems to be a result of the GB Esso promotional campaign to obtain 100 Tesco Voucher. Although this promotional campaign continues much time after the announcement, the sentiment in the 'tesco' dataset turned into more neutral, 67.1%, due to the penetration of the news about the Tesco & Booker merge to the general Twitter audience. Given that this information is for the general public more factual than sentimental, neutrality increases just before the announcement. The analysis of the $TSCO dataset is the reverse way, tweets under the $TSCO umbrella are usually factual or neutral, or at least neutral for a general polarity classifier, but when an impacting financial event occurs neutrality vanishes and a bi-polarity emerges in response to the merge. Some example tweets and their corresponding polarity are provided for illustrative purposes only in Table 8. To conclude, and with the precautions mentioned, we can establish the fulfilment of **RQ2** so we can perceive change in sentiment related to the financial event through time.

9 Geographical impact of the action

Although Twitter is one of the most used data sources in data mining, the geo-location component of Twitter is not comparable to other data sources which we can refer to as Location-based social networks. In fact, according to [33], the geo-located tweets returned by the Streaming API cover up to 90% of the geo-located tweets extracted from Firehose API. However, this study also reveals that the number of geo-located tweets is low, being only a 1.45% of the tweets obtained from Firehose API and 3.17% of the tweets obtained from Streaming API. The total percentage of Tweets geo-located in the 'tesco' dataset is consistent with this previous study (Morstatter, Pfeffer, Liu, & Carley, 2013), with a percentage of 4.3% for all the periods in the experiment. Although the number of tweets in the $TSCO dataset may be not representative enough, we should remark that the percentage of geo-located tweets in the $TSCO dataset is 0%, 1 tweet out of a total of 199. Also,

Table 7 Sentiment Analysis results for the 'tesco' and $TSCO dataset

	Pre-announcement (25-27 Jan 06:59 am)				Announcement (27 07:00am-29 Jan)			
	TUESDAY-WEDNESDAY-THURSDAY				FRIDAY-SATURDAY-SUNDAY			
	'tesco'		$TSCO		'tesco'		$TSCO	
	Tweets	%	Tweets	%	Tweets	%	Tweets	%
Positive	181	55.20%	1	12.50%	599	24.60%	30	33.00%
Neutral	90	27.40%	7	87.50%	1636	67.10%	43	47.30%
Negative	57	17.40%	0	0.00%	202	8.30%	18	19.80%
Total	328		8		2437		91	

Table 8 Example of tweets in 'tesco' and $TSCO' dataset and their corresponding polarities

Polarity	Tesco dataset	$TSCO
Positive	Tesco merging with Londis ?? f****n love Londis	Good chat with Steve Fox,Booker/Tesco merger is great news for independents, better access to banking/payment services/
Negative	Tesco PLC acquires Booker Group for 3.7B creating the "UK's leading food business" uniting the power of the UK's la	The Booker merger means that Londis and Budgens now under the Tesco umbrella.
Neutral	Tesco buying Booker Group - disastrous for convenience stores who buy stock from Booker Cash&Carry at prices competitive to supe	Tesco & Booker merger (aka takeover): potentially terrible news for independent #retailers & small grocery chains. Stakeholders beware!

there is not variability of those percentages throughout the periods considered (pre- post- and regular). Beyond the percentage of geo-located tweets that the Twitter APIs return, the variation of the geographical distribution of the tweets due to the financial event deserves to be analysed. Figure 10 shows this distribution and illustrates that there is not much variation if we compare post-announcement with the regular period the same days of the week (10th Feb 07:00 - 12th Feb 23:59 as defined in Table 1.

A deeper inspection of the tweets per country in Table 9 confirms that most of tweets come from the countries where Tesco PLC deploy its main business either under Tesco trademark or thorough subsidiary local companies. Apart from UK and Republic of Ireland, the main retail locations of Tesco PLC all over the world are Czech Republic, Hungary, Poland, Slovakia, Turkey, Malaysia and Thailand. According to the results in Table 9, before the announcement, the bigger contribution to Twitter volume corresponds to the UK market which is consistent with the historical roots of the company where its retailing business is fully integrated in the society. Nevertheless, after the announcement, this percentage decreases in favour of other locations over the world, which is a sign of the global impact of the merge so that twitter users outside UK are not so linked to Tesco PLC marketing campaigns during regular period but they are reactive to a relevant event related with a company with presence in their countries. Nigeria is highlighted in Table 9 as a country with a definitely high position according to the number of tweets during the post-announcement despite the fact that Tesco does not having presence in this country. 42 of the 43 tweets in Nigeria has the same

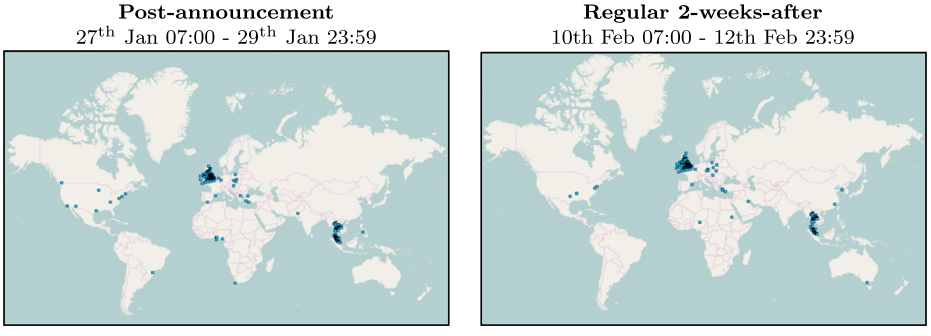


Fig. 10 Geographical distribution of tweets in 'tesco' dataset after the announcement and during a regular period

Table 9 Geographical distribution of tweets in 'tesco' dataset

Pre-Announcement			Post-Announcement			2 Weeks after (8th - 10th Feb)			2 weeks after (10th- 12th Feb)		
Country	Geolocated Tweets	%	Country	Geolocated Tweets	%	Country	Geolocated Tweets	%	Country	Geolocated Tweets	%
UK	443	61.27%	UK	585	51.32%	UK	333	58.12%	UK	503	55.76%
Malysia	196	27.11%	Malysia	341	29.91%	Malysia	162	28.27%	Malysia	242	26.83%
Thailand	53	7.33%	Thailand	125	10.96%	Thailand	33	5.76%	Thailand	100	11.09%
			Nigeria	42	3.68%	Ireland	18	3.14%	Ireland	22	2.44%
			Ireland	15	1.32%						
Rest of the world	31	4.29%		32	2.81%		27	4.71%		35	3.88%
TOTAL	723		TOTAL	1140		TOTAL	573		TOTAL	902	

content but they are tweeted from 42 different users, not being retweets, so that it may be a violation of the spam terms in Twitter rules. With all the above information, we can establish the fulfilment of the **RQ2**, so that the event impact on Twitter depends on the location and, moreover, the event distorts the geographical distribution of tweets.

10 Discussion & limitations

Coming back to the fundamental research question addressed in this paper, we sustain and successfully confirmed that "Twitter (although not a specific financial forum) is permeable to financial events and this permeability can be analysed by monitoring (1) the name of companies as a keyword, (2) the Cashtag of the company and (3) the hashtags related to that company." The results of the experiment on the announcement about the merge of Tesco PLC and Booker Group PLC on the 27th January 2017 show that the *Twittesphere* is permeable to the financial market dynamics thanks to the tweets of a variety of different contributors. The merge impacted on all the Tesco-related content in Twitter in terms of volume, having a higher impact on $cashtag-content but is altering the tweets' topics in comparison to regular behaviour (altering #hashtag dynamics and tweets' content). Also, with the precautions of a single experiment, we also observed changes in polarity and in the geographical distribution of the contributes through time. Finally, the good freshness characteristics of Twitter as news media is confirmed by the rapidness of response to the RNS announcement.

Therefore, the experiment was successful in confirming that a far-impacting financial event causes disturbance in all the features considered in relation with Twitter permeability: information volume, content and sentiment as well as geographical provenance. Nevertheless, the experiment had a little success in identifying some rumour or sign of the announcement prior to the event. Even considering that the experiment was not deployed over the whole Firehose Twitter data, uncovering rumours before the announcement turns definitively into a hard task, if the spreading of rumours in real life is not mimicked inside Twitter, that means, if the rumour is not in Twitter at all. At this respect, and according to [30], social media data can only be generalised to human behaviour when social media provides a representative description of human activity. Twitter is a social media which, at least, exhibits some demographic bias. Moreover, Twitter may be providing a skewed representation of content. Although well-known rumour detection algorithms [51, 53] can be applied to Twitter, an alternative approach can be the fusion of financial information from different data sources in a way that we can mitigate the inevitable bias in a single source, and, at the same time, combine their weaknesses and strengthens in a proper representation of the real financial activity.

As mentioned, the experiment in this paper addresses a well known financial market event as a first step to adapt the methodology to low impacting events. Presumably, financial events with fully coverage in the media (social media, news, tv) would have a major impact on the whole Twittersphere, meanwhile financial events in sectors not so attractive to

the general public as retailing or, not so covered by non financial media, would have lower impact on Twittersphere but remain impacting inside Twitter financial sphere. The permeability and patterns of change in hashtag and cashtag should be adjusted according to theses conditions. The real deployment of a continuous real-time monitoring system for irregularities coming from financial actions will be based on mining social media and establishing patterns for the time series of cashtag and hashtags and topics (bag-of-words) so we can apply and ARIMA (Autoregressive moving average model) to avoid noise and distinguish the stationary behaviour form the tending one. The continuous adjustments of the patterns, the bag of words and outlier detection techniques will be easily scalable to the huge quantity of data by a map-reduce implementation and a cloud deployment.

11 Conclusions

This paper inspects the permeability of Twitter to financial events in order to provide evidence which allows Twitter to be used as a social sensor for the financial and stock market. To do that, this permeability should be checked and measured. This a single experiment for a single financial that had been fully covered by traditional media as well as social media. Bearing this in mind, we can conclude that the event in the financial market invaded the Twittersphere on the 27th January 2017, just after the RNS announcement at 7:00, and that the behaviour of the triplet ($TSCO, #[tesco], "tesco") was altered in comparison with the regular behaviour around the company involved in the financial event. At the same time, the experiment shows that other financial events that affected the company (overstatement) during the very same period were only permeable to the more financial oriented tweets and users. As many other social networks, homophily [14] exists in the context of Twitter, that is, the tendency to interact with similar individuals in respect to several dimensions such as age, location, and occupation, etc. The experiment in this paper shows that the high impact of the merge action crosses the fragile boundary between users with specific financial interests (highly connected to each other) to the general audience in Twitter. However, this work has not studied the profile of the users in general audience that are captured by the Tesco merge, or, to put in another way, to which kind of homophilic words the information flows out from the experts in the stock market.

As mentioned before, this Twitter study is framed by a a joint project that pursue the construction of a Data Fusion Model and an Ecosystem to monitor the financial market, which is increasingly demanded given the irruption of FinTech technologies and the subsequent increase of non expert traders, automatic trading systems, cryptocurrencies, HFT, etc. The lack of integration between financial stock market data, social media comments, financial discussion board posts and broker agencies means that the benefits of data fusion are not being realised to their full potential. Our proposed ecosystem [20], inspired by the data fusion model introduced by JDL (Joint Directors of Laboratories [52]), pursue the fusing of disparate data sources relating to financial stocks; data with a diverse set of features from different data sources will supplement each other in order to obtain a Smart Data Layer, which will assist in scenarios of irregularity detection.

Acknowledgments This work was funded by Spanish Ministry of Education Culture and Sports, National Plan for Scientific and Technical Research and Innovation (Sub-Programme for Mobility) under the research stay grant PRIX16/00368. We thank the Manchester Metropolitan University (School of Computing Mathematics and Digital Technology) for its support during the research stay. This work is also partially funded by the Spanish Ministry of Economy and Competitiveness under the National Science Program (TEC2014-54335-C4-3-R, TEC2017-84197-C4-2-R).

References

1. Agarwal S, Chomsisengphe S, Cheryl L (2017) Consumer choice and financial products. Annu Rev Financ Econ 9:127–146
2. Al Nasseri A, Tucker A, de Cesare S (2014) Big data analysis of stocktwits to predict sentiments in the stock market. In: Džeroski S, Panov P, Kocev D, Todorovski L (eds) Discovery science. Springer International Publishing, Cham, pp 13–24
3. Antweiler W, Frank M (2004) Is all that talk just noise? The information content of internet stock message boards. J Financ 59(3):48
4. Azar P, Lo AW (2016) The wisdom of twitter crowds: predicting stock market reactions to fomc meetings via twitter feeds, available at SSRN: https://ssrn.com/abstract=2756815. https://doi.org/10.2139/ssrn.2756815
5. Billett M, Yu M (2016) Asymmetric information, financial reporting, and open-market share repurchases. J Financ Quant Anal 51(4):1165–1192
6. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
7. Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I (2012) Web search queries can predict stock market volumes. PLOS ONE 7(7):1–17
8. Bormetti G, Calcagnile LM, Treccani M, Corsi F, Marmi S, Lillo F (2015) Modelling systemic price cojumps with hawkes factor models. Quant Finan 15(7):1137–1156
9. Campbell J, Cecez-Kecmanovic D (2011) Communicative practices in an online financial forum during abnormal stock market behavior. Inf Manag 48(1):37–52
10. Cavalcante RC, Brasileiro RC, Souza VL, Nobrega JP, Oliveira AL (2016) Computational intelligence and financial markets. Expert Syst Appl 55(C):194–211
11. Cazzoli L, Sharma R, Treccani M, Lillo F (2016) A large scale study to understand the relation between twitter and financial market. In: 2016 third European network intelligence conference (ENIC), pp 98–105
12. Ceccarelli D, Nidito F, Osborne M (2016) Ranking financial tweets. In: ACM (edn) proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval (SIGIR '16), pp 527–528
13. Cortez P, Oliveira N, Ferreira JP (2016) Measuring user influence in financial microblogs: experiments using stocktwits data. In: ACM (ed) WIMS'16 proceedings of the 6th international conference on web intelligence, mining and semantics
14. De Choudhury M (2011) Tie formation on twitter: Homophily and structure of egocentric networks. In: 2011 IEEE third international conference on IEEE (ed) privacy, security, risk and trust (PASSAT) and 2011 IEEE 3rd international conference on social computing (SocialCom)
15. Delort JY, Arunasalam B, Leung H, Milosavljevic M (2012) The impact of manipulation in internet stock message boards. Int J Banking Finan 8(4):19
16. Dickinson BW (2015) Sentiment analysis of investor opinions on twitter. Soc Netw 4:62–71
17. Ding X, Zhang Y, Liu T, Duan J (2015) Deep learning for event-driven stock prediction. vol 2015-January
18. Dredze M, Kambadur P, Kazantsev G, Mann G, Osborne M (2016) How twitter is changing the nature of financial news discovery. In: ACM (ed) proceedings of the second international workshop on data science for macro-modeling
19. Elliott WB, Grant SM, Hodge FD (2017) Investor reaction to firm or ceo use of twitter for negative disclosures. SSRN
20. Evans L, Owda M, Crockett K, Fernández-Vilas A (2018) Big data fusion model for heterogeneous financial market data (findf). In: Intelligent systems conference. IntelliSys 2018
21. Fernández-Vilas A, Evans L, Owda M, Díaz Redondo RP, Crockett K (2017) Experiment for analysing the impact of financial events on twitter. Springer International Publishing, Cham, pp 407–419

22. Gunduz H, Cataltepe Z (2015) Borsa istanbul (bist) daily prediction using financial news and balanced feature selection. Expert Syst Appl 42(22):9001–9011
23. Hentschel M, Alonso O (2014) Follow the money: a study of cashtags on twitter. First Monday 19(8)
24. Hobijn B, Jovanovic B (2001) The information technology revolution and the stock market: evidence. Am Econ Rev 91:1203–1220
25. Hu T, Tripathi A (2016) Impact of social media and news media on financial markets. SSRN
26. Karppi T, Crawford K (2016) Social media, financial algorithms and the hack crash. Theory, Culture & Society 33(1):73–92
27. Li Q, Wang J, Wang F, Li P, Liu L, Chen Y (2017) The role of social sentiment in stock markets: a view from joint effects of multiple information sources. Multimed Tools Appl 76(10):12:315–12:345
28. Liew JKS, Budavári T (2016) Do tweet sentiments still predict the stock market? SSRN
29. Liu L, Wu J, Li P, Li Q (2015) A social-media-based approach to predicting stock comovement. Expert Syst Appl 42(8):3893–3901
30. Liu H, Morstatter F, Tang J, Zafarani R (2016) The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. Int J Data Sci Analytics 1(3-4):137–143
31. Loria S (2014) Textblob: simplified text processing
32. Miller GS, Skinner DJ (2015) The evolving disclosure landscape: how changes in technology, the media, and capital markets are affecting disclosure. J Account Res 53(2):221–239
33. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose. In: Proceedings of the 7th international conference on weblogs and social media, ICWSM 2013. AAAI Press, pp 400–408
34. Muhammad A, Leak A, Longley P (2014) A geocomputational analysis of twitter activity around different world cities. Inf Sci 17(3):145–152
35. Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL (2015) Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. Expert Syst Appl 42(1):306–324
36. Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42(24):9603–9611
37. Oliveira N, Cortez P, Areal N (2016) Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decision Support Syst 85:62–73
38. Oliveira N, Cortez P, Areal N (2017) The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst Appl 73(Complete):125–144
39. Owda M, Crockett K, Lee P (2017) Financial discussion boards irregularities detection system (fdbs-ids) using information extraction. In: Intelligent systems conference 2017
40. Pagolu VS, Reddy KN, Panda G, Majhi B (2016) Sentiment analysis of twitter data for predicting stock market movements. In: 2016 international conference on signal processing, communication, power and embedded system (SCOPES), pp 1345–1350
41. Rajesh N, Gandy L (2016) Cashtagnn: Using sentiment of tweets with cashtags to predict stock market prices In: 11th international conference on intelligent systems: theories and applications, SITA. IEEE
42. Ranco G, Aleksovski D, Caldarelli G, Grcar M, Mozetic I (2015) The effects of twitter sentiment on stock price returns. PloS one 10(9):e0138441
43. Ranco G, Bordino I, Bormetti G, Caldarelli G, Lillo F, Treccani M (2016) Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. PLOS ONE 11(1):1–14
44. Rao T, Srivastava S (2014) Twitter sentiment analysis: How to hedge your bets in the stock markets. Springer International Publishing, Cham, pp 227–247
45. Rodríguez-Domínguez D, Redondo RPD, Vilas AF, Khalifa MB (2017) Sensing the city with instagram: Clustering geolocated data for outlier detection. Expert Syst Appl 78:319–333
46. Ruiz EJ, Hristidis V, Castillo C, Gionis A, Jaimes A (2012) Correlating financial time series with micro-blogging activity. In: Proceedings of the 5th ACM international conference on web search and data mining, WSDM '12. ACM, New York, pp 513–522
47. Sabherwal S, Sarkar S, Zhang Y (2011) Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. J Bus Finance Account 38:1209–1237
48. Servia-Rodríguez S, Díaz-Redondo R, Fernández-Vilas A (2015) Are tweets biased by audience? An analysis from the view of topic diversity. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer International Publishing
49. Shutes K, McGrath K, Lis P, Riegler R (2016) Twitter and the us stock market: The influence of micro. bloggers on share prices. Econ Bus Rev 2(3):57–77
50. Sprenger TO, Tumasjan A, Sandner PG, Welpe IM (2014) Tweets and trades: the information content of stock microblogs. Eur Finan Manag 20:926–957

51. Tafti A, Zotti R, Jank W (2016) Real-time diffusion of information on twitter and the financial markets. PLoS ONE 11(8):e0159226
52. Välja M, Korman M, Lagerström R, Franke U, Ekstedt M (2016) Automated architecture modeling for enterprise technology manageme using principles from data fusion: A security analysis case. In: 2016 Portland international conference on management of engineering and technology (PICMET), pp 14–22
53. Vosoughi S (2015) Automatic detection and verification of rumors on twitter
54. Wang B, Huang H, Wang X (2012) A novel text mining approach to financial time series forecasting. Neurocomputing 83:136–145
55. Wu L, Hoi SC, Yu N (2010) Semantics-preserving bag-of-words models and applications. IEEE Trans Image Process 19(7):1908–1920
56. Xiong F, MacKenzie K (2015) The business use of twitter by australian listed companies. The J Developing Areas 49(6):421–428
57. Xiong F, Prasad A, Chapple L (2016) The economic consequences of corporate financial reporting on twitter. In: 7th conference on financial markets and corporate governance conference
58. Zhang L (2013) entiment analysis on twitter with stock price and significant keyword correlation. PhD thesis, University of Texas
59. Zheludev I, Smith R, Aste T (2014) When can social media lead financial markets? Sci Report 4:4213

Ana Fernández Vilas I am Associate Professor at the Department of Telematics Engineering of the University of Vigo and researcher in the Information & Computing Laboratory (AtlantTIC Research Center). I received my PhD in Computer Science from the University of Vigo in 2002. My research activity at I&C lab focuses on Semantic-Social Intelligence & data mining. I look for applying both to Ubiquitous Computing and Sensor Web; urban planning & learning analytics. Also, I am involved in several mobility & cooperation projects with North African countries & Western Balkans.

Rebeca P. Díaz Redondo is an Associate Professor at the Department of Telematics Engineering of the University of Vigo and researcher in the Information & Computing Laboratory (AtlantTIC Research Center). She received her PhD in Telecommunications Engineering from the same university. She currently works on applying social mining and data analysis techniques to characterize the behavior of users and communities to design solutions in learning, smart cities and business areas. She is currently involved in the scientific and technical activities of several national and European research & educative projects. Besides, she is involved in several mobility & cooperation projects with North African countries & Western Balkans.

Keeley Crockett is Reader in Computational Intelligence in the School of Computing, Mathematics and Digital Technology at Manchester Metropolitan University and a Senior Fellow of the Higher Education Academy. She gained a BSc Degree (Hons) in Computation from UMIST in 1993, and a PhD in the field of machine learning from the Manchester Metropolitan University in 1998 entitled Fuzzy Rule Induction from Data Domains. She obtained a P.G.C.E from The University of Huddersfield in June 2000. She also holds an Institute of Line Management Level 5 Diploma for Professional Management Coaches and Mentors. Her main research interests include the areas of fuzzy decision trees, rule induction, applications of fuzzy theory, conversational agents including intelligent tutoring systems, psychological profiling with neural networks Silent Talker and FATHOM), short text semantic similarity and data mining Big Data. She is author or co-author of more than 84 papers published in specialized journals and congress proceedings. She has also adapted and co-authored two international books on Database Principles. She is currently Co-Leader of Computational Intelligence & Reasoning (CIR) and was former leader of the Intelligent Systems Group, which has established a strong international presence in its research into Conversational Agents and Adaptive Psychological Profiling including an international patent on Silent Talker. . Nationally, she is the current (2015 -) Chair of IEEE Women in Engineering for the United Kingdom and Ireland. She is the current Chair of the IEEE Computational Intelligence Society Student Activities committee (2016), Vice-chair of the IEEE Computational Intelligence Pre-college Activities committee, Vice Chair (2014-) of the IEEE Women into Computational Intelligence Sub-committee (was Chair 201s2-2014),

Majdi Owda is a senior lecturer in Computer Science in the School of Computing, Mathematics and Digital Technology at Manchester Metropolitan University. He gained a BSc in Computer Science from the Arab American University in 2004, and an MSc by research in Computer Science with distinction from Manchester Metropolitan University in 2005 and a PhD in Computer Science in 2012. During the BSc Majdi specialised in Software Engineering and Web Applications Development. In the MSc, he concentrated on Artificial Intelligence with a thesis titled "Cased-Based Reasoning and Pattern Matching for Automatic Email Response". In his PhD research he worked on the creation of Conversation-Based Interfaces to Relational Databases (C-BIRDs) through the use of Conversational Agents and AI techniques. His main research interests are Natural Language Interfaces to Relational Databases, Conversational Informatics, Conversational Agents, Knowledge Trees, Knowledge Engineering, Information Extraction, AI techniques for crime prevention and Web/Data/Text Mining.

Lewis Evans is currently pursuing a PhD at the School of Computing, Mathematics and Digital Technology at Manchester Metropolitan University. Mr Evans graduated with a first-class honours degree at Manchester Metropolitan University in 2015; BSc Computer Forensics and Security. Mr Evans' PhD is titled "A Smart Data Ecosystem for Continuous Monitoring of Financial Market Irregularities".