UK Ph.D. Centre in Financial Computing

UNIVERSITY COLLEGE LONDON

Ph.D. in Financial Computing

Thesis

When Can Social Media Lead Financial Markets?

Ilya Nikolay Zheludev

Supervisor: Dr. Robert Smith

Secondary Supervisor: Dr. Tomaso Aste

March 2015

This Thesis is submitted as part requirement for the Ph.D. degree in Financial Computing at University College London.

I, Ilya Zheludev, confirm that the work presented in this Thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the Thesis.

Ilya Nikolay Zheludev

ABSTRACT

Social media analytics is showing promise for the prediction of financial markets. The research presented here employs linear regression analysis and information theory analysis techniques to measure the extent to which social media data is a predictor of the future returns of stock-exchange traded financial assets.

Two hypotheses are proposed which investigate if the measurement of social media data in real-time can be used to pre-empt – or *lead* – changes in the prices of financial markets. Using Twitter as the social media data source, this study firstly investigates if geographically-filtered Tweets can lead the returns of UK and US stock indices. Next, the study considers if string-filtered Tweets can lead the returns of currency pairs and the securities of individual publically-traded companies. The study evaluates Tweet message *sentiments* – mathematical quantifications of text strings' moods – and Tweet message *volumes*. A sentiment classification system specifically designed and validated in literature to accurately rank social media's colloquial vernacular is employed. This research builds on previous studies which either use sentiment analysis techniques not geared for such text, or which instead only consider social media message volumes. Stringent tests for statistical-significance are employed.

Tweets on twenty-eight financial instruments were collected over three months – a period chosen to minimise the effect of the economic cycle in the time-series whilst encapsulating a range of market conditions, and during which no major product changes were made to Twitter. The study shows that Tweet message sentiments contain lead-time information about the future returns of twelve of these securities, in excess of what is achievable via the analysis of Twitter message volumes. The study's results are found to be robust against modification in analysis parameters, and that additional insight about market returns can be gained from social media data sentiment analytics under particular parameter variations.

ACKNOWLEDGEMENTS

This juncture would have been unreachable without my parents’ unwavering generosity, love and boundless support – for this I am deeply indebted to them. Tanya Nousinova and Nikolay Zheludev, I dedicate this Thesis to you, and will always use you as examples of how to give one’s children every possible opportunity. I also wish to thank my brother Vanya, my grandmother Galina and my girlfriend Emily, for their patience, understanding and kindness throughout this journey.

To my resilient first supervisor, Dr. Robert Elliott Smith, I offer my deepest thanks. He took me under his wing, and offered me unparalleled and undeniably selfless guidance throughout this research and writing. Without him, none of this would have been possible. Robert, you are undoubtedly the benchmark, and I consider having your supervision as a gift I should have never deserved.

To my dedicated second supervisor, Dr. Tomaso Aste, I offer my humble gratitude. His altruistic support and perseverance when all else had failed gave me the energy to see my research through to the end. Tomaso, your effortless wisdom was instrumental to the success of this work.

I also wish to thank my dear friends, the Goodeal family. Your peaceful and caring attitude helped me wind down when my work became exhausting. Hemish, our friendship and your advice over the years was invaluable in helping me complete my research – for this I will always be grateful. Niall, you asked for nothing, but were always the voice of reason & logic – and I thank you for this. Chaaya and Hem, your loving home and caring dispositions are the marks of comfort – I hope our friendship will thrive indefinitely.

Finally, to Sam Armstrong, I thank you for your steadfast support and precious guidance during the darkest hours. May your future bring you joy and prosperity, and may our friendship continue for many more decades.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF LISTS

1 INTRODUCTION

This chapter presents an overview of the problem addressed in this study. A brief background to the area of research is provided. Hypotheses and aims are stated.

1.1 Research background, motivation and context

The proliferation of the internet into every aspect of our lives has improved our ability to access data in real time. The internet has evolved over the last thirty years into a source of information on almost any topic or even thought. The US Department of Defence's adoption of the TCP/IP internet Protocol Suite as the standard for all military networking in 1983[1] was followed by the invention of the World Wide Web[a] by Berners-Lee in 1990. By 2012, this technology was routinely used by over 38% of the world's population[2] for both professional and leisure purposes.

A particular implementation of the internet that has seen growth in the 21st century[3] is social media[b], an example of which is Twitter[c], a micro-blogging and personal-message sharing service started in 2006, and floated on the New York Stock Exchange in 2013[4]. The company, which handles over 500 million users and over 500 million daily messages, is used globally by a broad demographic[5] to publically broadcast, or 'Tweet' 140-character messages on any chosen topic. This allows internet users to broadcast their thoughts to a global audience in real time and at zero cost. The implications are that for the first time in human history, it is possible to monitor the moods, thoughts and opinions of the world's population in an aggregated and real-time manner with almost negligible data-collection costs. Social media data have been used to predict real-world phenomena such as brand popularity[6], silver-screen box office returns[7] and election outcomes[8]. Of present focus is the prediction of financial markets via the analysis of Tweets[9-12] and other comparable data sources such as Google[d] Search Trends[13-15], Yahoo![e] search engine data[16] and Wikipedia[f] articles[17]. Whilst the rationales behind these analyses are united together by the suggestion of information inefficiency in

[a] A series of interlinked, remotely-stored hypertext documents accessible via the internet.

[b] Interactive internet-based platforms via which individuals and communities create and share user-generated content.

[c] http://www.twitter.com

[d] A web search engine owned by Google, Inc. It is the world's most-visited website.

[e] A web search engine owned by Yahoo, Inc. It is the world's fourth most visited website.

[f] A collaboratively-edited internet encyclopaedia, owned by the non-profit Wikipedia Foundation. It is the world's sixth most-visited website.

financial markets[18,19], there are still questions on the effectiveness of these potential predictive indictors, stemming from the use of *ex-post facto* techniques to structure portfolios retroactively with profit maximisation as the success criterion. We are still far from a unified consensus the extent to which we can anticipate the financial markets with social media data, and we are still unaware of the quantity of information that social media contains on the future returns of market-traded securities.

This new area of research is in its dawn; is computationally challenging due to the size of the datasets that require analysing; and is at times open to scepticism. A multitude of questions remain unanswered, specifically relating to how textual information from the internet can be translated into predictive indicators, and whether or not there are indeed any predictive powers contained within. Differing data analysis approaches, both commercial and academic, have shown various levels of success, with the topic of Twitter leading the markets seeing discussion in the press[20-23]. Furthermore, since Twitter's user base is not an accurate sample of the world's population[5], it is reasonable for one to doubt the capacity of Tweets to accurately lead the performance of financial markets ahead of time, even without delving into the mathematics of the problem.

Of particular interest is the possibility of structuring investment portfolios based on signals from social media, and thus giving credence to previous such attempts. However, this study does not describe a trading strategy, nor is it a predictive-indicator generator. Rather, it is a necessary and currently-overlooked precursor to validate with stringent tests for statistical-significance, the extent to which social media data can contain *ex-ante* information about the future returns of market-traded securities, achieved without any biases associated with profit-maximisation or portfolio-structuring from such data.

1.2 Problem statement

This study centres on assessing if the analytics of social media data can be used to pre-empt – or lead – changes in the prices of financial markets. Given the availability of a randomised sample of all Tweets from a particular country, is it possible to lead the returns of that country's major stock indices? The rationale here is that the quantitative mood of aggregated Twitter messages from a particular country may act as an approximation of that nation's overall well-being, which in turn could correlate with, or lead its major stock indices. Furthermore, if Tweets are filtered by specific traded instrument identifiers and/or company names, can the quantitative mood of aggregated Twitter messages lead the returns of individual market-traded securities?

The scope of this study is to ascertain if social media data contains powers to lead the returns of market-traded assets and to what extent, without any bias associated with structuring investment portfolios from such signals. The results of this investigation would give further credence to the possibility of designing profitable trading strategies based on the analytics of social media data – however this is a separate exercise altogether; is currently premature and is therefore outside the scope of this study. Recent academic exercises[13-15] in portfolio structuring based on the analysis of social media data have placed particular emphasis on the analytics of message volumes, with retroactively-calculated profit maximisation as the success criterion. These methodologies are arguably not demonstrators of social media's ability to lead the markets, but rather are exercises in optimum-parameter selection which also typically overlook a valuable additional data-source: the quantitative moods of social media messages. The rationale behind conducting the study is therefore to support the viability of portfolio-structuring endeavours such as the aforementioned, but without using retroactively-calculated profit maximisation as a factor of the research. Instead, the study seeks to identify the value the quantitative moods – or sentiments – of social media messages for the profit-seeking investor. The present research therefore aims to improve our understanding of which subsets of social media data are of greater importance to investment practitioners, and thus to provide support for past social media-based portfolio-structuring initiatives.

1.3 Objectives and hypotheses

The following two hypotheses are evaluated in the study:

Hypothesis One: “The analysis of randomised samples of 10% of all Tweets from the United States and the United Kingdom can be used to lead the returns of S&P500 and FTSE100 indices, respectively”.

Hypothesis Two: “The analysis of Tweets filtered by instrument identifiers[a] and/or company names can be used to lead the returns of market-traded securities”.

FIGURE 1: STUDY HYPOTHESES

The hypotheses listed in Figure 1 are tested both by the quantification of the moods of Twitter messages, and also by the analysis of the volumes of the Tweets relevant to each case. Hence, it is established if the moods of Tweets carry additional powers to lead the markets, over and above what is attainable with the evaluation of just Tweet message volumes.

The study’s experiments are performed without the use of retroactive profit-maximisation as the success criterion or without structuring a trading strategy. Instead, this study answers the question of whether social media data contains statistically-significant information about the future returns of financial markets.

This Thesis covers: the problem at hand; the methodologies behind the collection of Twitter data; past initiatives into this area of research with a specific emphasis on the limitations of leading the markets using social media message volumes; the advantages of the quantitative analysis of the moods of Tweets and the optimum methodologies herein; the analytics involved in determining if a time-series of social media data can lead a time-series of financial data; the results of the investigation; the robustness of the results against parameter variation and an insight into the commonalities, drivers & generalisations of the findings.

[a] This is an abbreviation used in the financial services industry to uniquely identify publicly traded financial instruments such as shares of a particular stock on a particular stock-market.

1.3.1 Intended contributions of the study

The study's intended contributions are:

- An improved understanding of the real value of this new data-source for use as a variable for leading the markets, ascertained without portfolio-structuring bias and without retroactive profit-maximisation as the success criterion as is the case with recent studies[13-15]. This is instead achieved by the quantification of the amount of information that Twitter data contains about the returns of market-traded securities ahead of time;

- A statistically-significant validation of whether Twitter data can lead the returns of individual market-traded companies and/or stock indices;

- An in-depth insight into the extent to which the quantitative moods of Tweets can lead the markets over and above what is available from the analytics of social media message volumes. This analysis reveals limitations in what can be expected from social media data in leading securities' returns ahead of time;

- An insight into the generalisations of the extent to which social media message volumes can be an indicator of message sentiment being able to lead the returns financial securities and to what extent;

- An insight into the generalisations of the extent to which message sentiment adds predictive powers to message volume when leading the returns of financial securities with social media data;

- The above are achieved via the creation of a series of data collection and analytics frameworks for connection to, and the evaluation of, Twitter data for the study;

- And finally, the conceptual design, management and construction of SocialSTORM – UCL's Social Media Analytics Engine. As part of the study, SocialSTORM was brought together from conception to realisation at the start of the study in order to provide UCL with access to social media data for research purposes (see Chapter 4.1). Data from SocialSTORM was used for preliminary experiments in the study.

2 BACKGROUND INFORMATION

This chapter presents a background to the study, with emphasis on technical aspects of the investigation.

The existence of social media is a revolution. For the first time in history, it is giving the everyday man or woman the ability to reach a potentially unlimited audience at an insignificantly low cost. This is revolutionary because social media is changing the broadcast-communication model. Before the existence of social media, communication & broadcasting to large and substantial audiences was restricted to professionals using the mediums of Radio, Television and Print. This is the traditional 'few to many model', which was heavily-dependent on infrastructure. A need existed for: cameras; microphones; studios; the printing press; directors; producers; photographers *et cetera.* Each one of these is prohibitively expensive for conveying a casual message. Furthermore, the direction of communication was mostly one way: professionals, to listeners. But today, the individual's capacity to reach an audience of thousands, if not millions is mostly dependent only on the quality of the message being conveyed. Furthermore, the direction of communication is now two-way – we are all simultaneously broadcasters and listeners[24]. Simply broadcasting interesting content can be the only criterion sufficient for an individual to reach a global audience.

One particularly popular service that is allowing its users to achieve this is Twitter[a]. It is an example of micro-blogging[b], and it has been shown that users broadcast information for two purposes: to micro-blog about themselves, or to disseminate/share information[25]. Either way, the type of the information being conveyed is of interest to both users of social media and researchers since it spans topics such as: users' current activities; conversations amongst friends; reaching out to community members; posting web-links; or real-time news reporting[26].

Part of this study is therefore concerned with the collection of data from this rich source of information.

[a] Accessible via www.twitter.com

[b] An internet-based broadcast method similar to blogging but using shorter messages. With Twitter, each message, or 'Tweet' as it is known, is under 140 characters in length.

2.1 Availability of Twitter data

Public data from Twitter can be acquired free of charge, and are typically accessed by querying its Application Programming Interface (API)[a]. This can be used to tailor results according to a desired dataset via proprietary code. Twitter allows developers to track up to 400 specified keywords for which to filter publicly available Tweets before streaming to the developer in near real-time. It is also possible to filter Twitter data by user ID or location, achievable with HTTP POST requests. Obtaining a random sample of data from Twitter is also simple; the following HTTP GET request returns a live stream of 1% of all public Tweets as a JSON[b] array:

https://stream.twitter.com/1/statuses/sample.json

Furthermore, elevated access to a random sample of 10% of all global Tweets can obtained for academic research purposes. However, once these Twitter data have been published and streamed through its API, the data cease to be accessible. This highlights the need for continuous communication with Twitter, and suitable technologies for storage of the data to allow aggregation of a substantial dataset over time.

As part of this study, access to 10% of Twitter's data was contractually secured. A copy of this contract is available in the Appendix of this Thesis (see Chapter 11.1).

2.2 Computational evaluation of Twitter data

Before analysing data from Twitter, one must extract the relevant data fields from their raw JSON format. Figure 2 provides an anonymised example of the information fields returned for each Tweet retrieved via Twitter's API.

[a] A programmatic interface which allows different software components or software systems to exchange information.

[b] JavaScript Object Notation: a standardised method for the programmatic exchange of text.

```
{
        "text":"If you buy the iPad mini…you’re crazy stupid",
        "entities":{"user_mentions":[],"hashtags":[],"urls":[]},
        "contributors": null,
        "place": null,
        "id_str": null,
        "coordinates": [removed],
        "source": null,
        "retweet_count": 0,
        "in_reply_to_user_id": null,
        "in_reply_to_status_id": null,
        "favorited": false,
        "geo": null,
        "in_reply_to_screen_name": null,
        "truncated": false,
        "in_reply_to_status_id_str": null,
        "user": [removed],
        "retweeted": false,
        "id": [removed],
        "in_reply_to_user_id_str": null,
        "created_at":"Tue Oct 23 18:11:16 +0000 2012"
}
```

FIGURE 2: EXAMPLE RESPONSE FROM TWITTER'S API

Twitter offers many forms of metadata which can also provide a source for analysis, as well as the Tweet text itself. Examples include: location tags and the number of times the message is ‘retweeted’ (re-shared by other users, thus increasing the message’s audience). Metadata may consist of: integers; strings; or a combination of both. In order to extract these data, one needs to parse the raw JSON structure and store each desired string or integer as a separate variable. The data can then, for example, be stored within separate columns of a database, or as a text file with a specified delimiter.

Text data can be analysed in a number of ways, from simple message volume analysis to using techniques from Natural Language Processing[a][27]. For example, one such method which offers a quantitative insight into the meaning of text-based data is sentiment analysis[b][28]. This study is centred on the analytics of the sentiment of text for the programmatic extraction of mood from Twitter’s data-streams to ascertain the additional value that social media message sentiment data has in leading the markets, over the analytics of social media message volumes.

[a] The application of Computer Science techniques concerned with interactions between human language and machine.

[b] The application of Natural Language Processing to classify the polarity of a given string with regards to emotional opposites, e.g., ‘happy’ vs. ‘sad’.

2.3 The application of social media data analysis to informational inefficiency in financial markets

The theory of informational efficiency in markets states that the price of a financial asset is the product of all the publically available information on that asset. It pertains to the idea that the price of an instrument, as set by market forces, takes into account all of the publically-available knowledge which can influence the demand for, supply of, and hence the price of that asset. This bold statement is frequently disputed, both in commercial and academic circles[18] for instance because there is publically available and valuable information about an asset which cannot be processed, analysed or viewed by everyone in the marketplace in a sufficiently timely manner, often for technical reasons[19].

Hypotheses exist which state that information is contained within internet-based news data which can further support the notion of informational-inefficiency in financial markets[29]. Since such inefficiencies are suggested to result in incorrect asset pricing by the financial markets, social media data can assist with discovering correct asset prices. Undoubtedly, social media data will contain both noise and signal if such information does indeed exist. If this signal is real and if it can be utilised in a timely manner conducive to practical trading, it can be used to predict the future returns of market-traded securities for profit-making purposes.

3 LITERATURE REVIEW

This chapter presents a background to the field of using internet data to measure the effect of real-world phenomena, as well as a critical assessment of past work in the field of market-monitoring with social media data.

3.1 Measuring and tracking real-world phenomena with social media

Real-time social data from user-contributed social media platforms such as Twitter and Facebook, as well as query volumes from search-engines are being used to track real-world phenomena across a broad range of specialisations, not exclusively relating to the financial markets. Real-time data from the social web provides the ability to observe public opinion and activity without the reporting lags associated with the production and release of any government-agency data on real-world phenomena[30]. Social data have been used to track, predict and measure: epidemiological variables[31,32]; economic variables such as unemployment levels[33], the demand for automobiles[30] and consumer consumption metrics[34]; the popularities and sales of video games, music tracks and feature films[35]; the happiness of internet users as a proxy for the happiness of nations[36]; and the outcomes of political races[37].

An important distinction which must be made is whether such data are being used to predict the future, or to track the present. The latter, known as “nowcasting”, aims to utilise social and internet-derived data to quantify real-world phenomena in real-time[38] and ahead of the releases of any government-agency data – an endeavour which has been used to track: the present-moment happiness of nations[39,40]; real-time mortality rates[41] and influenza outbreaks[42]; voting intentions during political races[43]; and live macroeconomic activity[44,45]. Nowcasting financial markets is by contrast, unnecessary: real-time markets data is readily available. It is therefore important to note that this study is focused on ascertaining the future performance of financial markets. There is little point in nowcasting asset prices that are readily available in real-time – and therefore the focus is on whether there is leading information about asset prices in social media data.

3.2 The demographics of social media users

We are at the dawn of using the analysis of social media data to track or lead the returns of market-traded securities. However before considering the methodologies involved in doing so, we should assess the demographics of social media users. Are these users an accurate representation of society? Indeed it is argued that that they are not. For example, the Twitter 'population'[a] is a highly non-uniform sample of the real population[5]. For the United States, it has been shown that with regards to gender, a bias exists amongst Twitter users towards males, which has diminished since Twitter's launch (90% of users in 2007 to 60% of users in 2009). Secondly, it has been shown that whilst the racial demographics of Twitter users are often representative of society, variations exist depending on their geographical locations within the United States. In comparison to actual population figures, Hispanic users are underrepresented in Twitter's population in the USA's South-West; African-American users are underrepresented in the South and Midwest; and Caucasian users are overrepresented in major cities. Finally, as it is impossible to extract the age of a Twitter user from their profile, or infer it in any reliable way, the user demographics of a comparable social network may be used as an educated approximation.

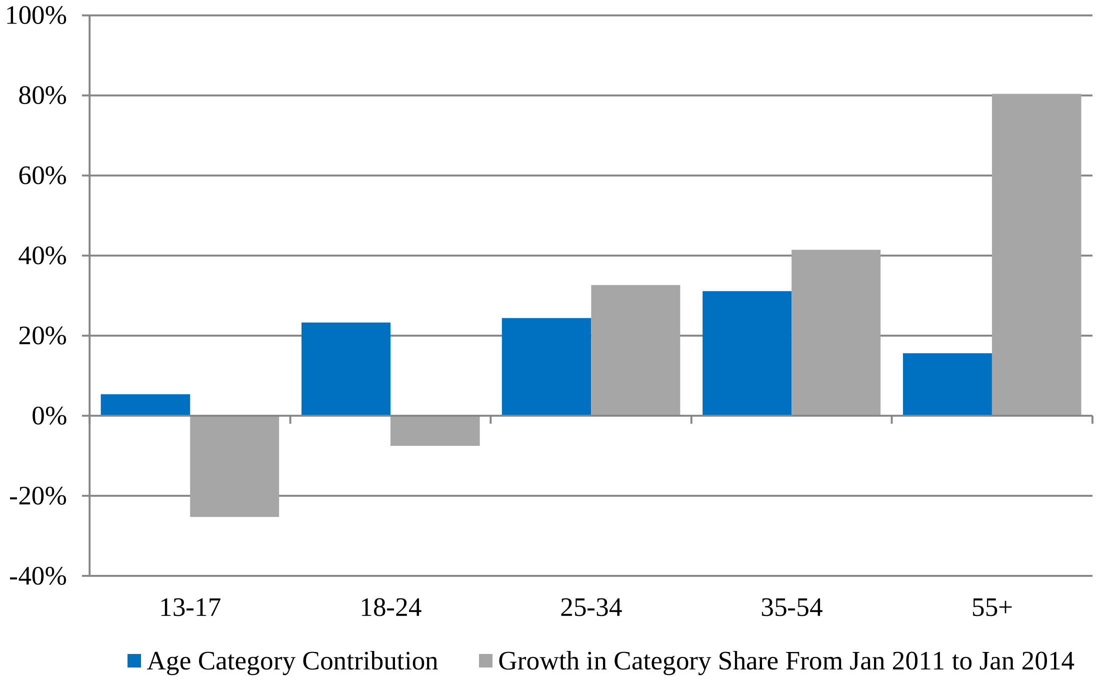


FIGURE 3: AGE DEMOGRAPHICS OF FACEBOOK USERS IN JANUARY 2014

[a] Defined here as Twitter's user-base, and its demographics.

Figure 3 shows the age of Facebook[a] users for January 2014, and the change from January 2011[46]. In particular, it should be noted that the largest age category by representation is 35-54, with the greatest category share increase from January 2011 being attributed to the 55+ age group.

3.3 Analysing social media and internet data to track or lead financial markets

There are at least two schools of thought regarding the best methodologies for tracking or leading financial markets using the analytics of text strings from social media. The first centres on the evaluation of changes in volumes of social media[11,12] messages, search engine queries[13-16] and Wikipedia article views and edits[17] to track and predict market movements, looking for statistically-significant relationships with the returns of stocks and indices. However, such studies do not quantitatively evaluate the content of social media messages, articles and search queries – a valuable source of data – and instead consider just their volumes[11-17]. The second methodology centres on attempts to track or lead financial market movements via the quantitative evaluation of the content of social media messages[9,10]. Such methodologies for anticipating markets ahead of time are typically performed via the concurrent quantitative analysis of the meaning of internet messages from large groups of individuals in advance of price changes in financial markets. When applied to the analysis of a group's thoughts on a particular topic, an average estimate from many individuals can offer stronger insights than the viewpoints of just the individual[47]. The computational analysis of the moods of social media messages is one way of ascertaining this "collective wisdom"[47,48] on a given topic. Known as sentiment analysis, the tool is a Natural Language Processing and Opinion Mining subtopic[49,50] which can allow for the classification of the polarity of unstructured text strings with regards to emotional scales, e.g., 'calm' vs. 'anxious'. The analysis of the sentiments of messages therefore allows for a deeper evaluation of social media's powers to lead financial markets, over and above what is possible with solely message-volume based analyses. However, the extent of the power of sentiment analysis methodologies in financial market prediction applications is still unknown, and is therefore the primary scope of this study.

[a] Facebook is a social networking service founded in 2004, and floated on the NASDAQ in May 2012. It is the world's second most-visited website.

The task of monitoring the sentiments of social media data has been considered since Twitter's launch in 2006. The application of sentiment analysis to unstructured and informal-vernacular internet-sourced text in particular is explored by Thelwall et al[50]. They recognise that a large number of currently-existing sentiment analysis tools are either not suitable for research purposes as the quantification methodologies are hidden and cannot be altered by users; or are not specifically designed to rank the colloquial nature of Twitter and internet-sourced text. Thelwall *et al.* thus present a research-oriented, transparent system known as SentiStrength[a] which outperforms baseline competitors in ranking the colloquial nature of user-generated text from internet platforms[50]. SentiStrength works on the principle of dictionary-matching, and is strongly based on the work of Pennebaker *et al.*[51], who created a multi-mood dictionary-term matching software called LIWC[b]. Alternative approaches are available – for example, "part of speech" tagging methodologies[52], but these are not specifically designed to accurately rank the often-informal and colloquial vernacular used on the internet as they rely on the standard spelling and grammar rules which are often ignored by social media users[36]. However, the implementation of such tools has shown that Twitter can be used as a measure of the world's happiness[30]. This is because the discussion of events in social, political, cultural and economic spheres does indeed take place on Twitter[26], with similar observations seen in web-search data[37].

Whilst the predictive analysis of Twitter messages has found a use in areas such as political tracking, for example in leading political election results[53] and characterising political debate performance[54], it is in finance that its use is of particular current interest. In this study, the notion of informational-inefficiency in financial markets[29] (as discussed in Chapter 2.3) is combined with the ability to monitor the moods of Tweets as outlined above, to assess if Twitter data can be used to lead the financial markets. Such research is underpinned by idea that a nation's mood is strongly linked to the performance of its stock indices, and vice-versa[55]. Early work[10] in the area of using social media analytics to lead stock indices showed that the mood of a random sample of 1% of all global Tweets significantly negatively correlated with the Dow Jones Industrial Average, NASDAQ and S&P500 indices, but significantly positively correlated with the Chicago Board Options Exchange Market Volatility index (VIX). As the first foray into this area of research, this work by Zhang *et al.* used a primitive 'part of speech' sentiment classification methodology that is not specifically designed for the

[a] http://sentistrength.wlv.ac.uk/

[b] Linguistic Inquiry and Word Count, http://www.liwc.net/

accurate ranking of the colloquial vernacular used in Tweets. Perhaps a more applicable approach is that of Bollen *et al.*[9] which addresses the same question by using a multi-mood approach. Employing a tool named GPOMS[a] which categorises text in six dimensions of emotion[b], it was shown that a random sample of global Tweets predicted the direction of the Dow Jones Industrial Average index with an accuracy of 86.7%. This research was then commercialised into an investment company named Derwent Capital Markets[c]. However, the hedge fund failed in 2010/2011[20], and has since been acquired (and subsequently dissolved) by an undisclosed financial technology firm[56]. These approaches were centred on the problem of predicting the future returns of stock indices, and not specific stocks. The study's first hypothesis[d] (see Chapter 1.3) addresses this issue by exploring the extent to which the analysis of Twitter messages filtered by country of origin can lead the returns of those countries' stock indices.

The issue of correlating or predicting the future performances of specific stocks with Twitter data, rather than just indices, is considered by Ruiz *et al.*[12]. As one of the first forays into answering this question, Ruiz *et al.* were only interested in correlations between message volumes on companies, and market trading volumes. Whilst this methodology cannot be used to predict asset returns, it demonstrated the availability of information in Tweets that could be exploited for market insight purposes. By monitoring Tweets mentioning industry-recognised company tickers[e], it was shown that Twitter data volumes significantly correlate with market trading volumes for certain publicly traded stocks. It was also shown that by using a linear regression model, the daily number of Tweets that mention S&P500 stocks significantly correlate with S&P500 daily closing prices, daily price changes and absolute daily price changes[11]. Similar observations have also been seen via comparable analyses of related data sources such as Google Search Trends[13-15], Yahoo! search engine data[16] and Wikipedia data[17], instead of Twitter data. Presently however, the availability of research on using sentiment analysis to lead the prices of specific stocks, rather than indices, is sparse. It is perhaps best attempted by Oliveira *et al.*[49], who provide a preliminary assessment of the content of Twitter data for identifying future performances of some specific stocks. However, their approach considered only nine stocks, with data amassed over only

[a] GPOMS: Google-Profile of Mood States.

[b] Calm, Alert, Sure, Vital, Kind and Happy.

[c] http://www.derwentcapitalmarkets.com/

[d] "The analysis of randomised samples of 10% of all Tweets from the United States and the United Kingdom can be used to lead the returns of S&P500 and FTSE100 indices, respectively".

[e] Tickers are reference codes used to denote different market-traded instruments specific to particular companies. "$AAPL", for example, relates to Apple, Inc. shares traded on the NASDAQ stock exchange.

thirty-two days. Furthermore, Oliveira *et al*. employed only simple regression analyses, and therefore leave open substantial scope for building on their evidence of returns predictability of specific stocks with Twitter sentiment data. The study's second hypothesis[a] (see Chapter 1.3) addresses this issue by exploring the extent to which the analysis of Twitter messages on specific publicly traded companies can lead the returns of their securities ahead of time.

It should also be noted that Twitter is not the only possible source of social media data which could be used to predict market movements. For example, queries from internet search engines could contain predictive powers – such sources are comparable to the data which can be extracted from Twitter. Whilst such data sources are outside of the scope of the this study, past work in this area has shown that search engine message volumes show correlations with financial trading volumes[15,16]. Such works have been extended further to structuring and trading based on the volumes of search engine queries. For example, profit-making trading strategies have been demonstrated based on the analysis of particular terms from Google Trends[13]. However, these strategies were structured based on an *ex-post facto* identification of the search terms which would result in the highest profits retroactively. These are therefore not necessarily demonstrations of social media or the internet's ability to lead financial markets – but are rather exercises in optimum-parameter identification for the purposes of maximising profits from datasets of historic financial data in a back-dated fashion. In fact, in such a manner it is possible to develop back-dated profit-making trading strategies based on the volume analysis of internet and social media terms that have no reference to economics or finance. One such study by Challet *et al.*[14] demonstrates that the volumes of internet searches relating to random non-finance terms such as: illness, cars, and arcade games, contains as much exploitable predictive information as finance-related terms, when considered in a similar *ex-post facto* parameter-selection configuration. Therefore, to ascertain social media or the internet's ability to lead financial markets without using *ex-post facto* methods, what is required is an analysis which measures the quantity of information contained within social media data on the returns of market-traded securities ahead of time. Here, rather than determining which terms could have generated profitable strategies retroactively, a more fundamental question is explored in this study: does social media data even contain enough statistically-significant information on the financial markets to act as a lead-time indicator? Furthermore, to

[a] "The analysis of Tweets filtered by instrument identifiers and/or company names can be used to lead the returns of market-traded securities".

what extent can social media sentiments offer additional abilities to lead financial returns over and above the analytics of social media message volumes? It is hoped that by answering these questions, support and further credence can be provided to this research space, in particular to past works which used *ex-post facto* identification of the most profitable search terms for retroactive-based portfolio-structuring.

4 ACCESSING TWITTER DATA

This chapter details the methodologies used to access data from Twitter.

One of the key technical drawbacks in conducting this study is the unavailability of aggregated datasets of historic Tweets. However, due to the nature of the License Agreement between Twitter and its users, most Tweets are in the public domain and can be accessed programmatically and then stored locally. However, due to the large volume of messages being passed through Twitter's API, a computational challenge exists in being able to store, access and analyse these Tweets in a timely manner.

To tackle this task, a cloud-based 'central-hub' was built during the start of this study. This platform facilitated the acquisition, storage and analysis of live data from various social media feeds. Known as SocialSTORM[a], the platform was a **St**reaming, **O**nline **R**epository and analytics **M**anager designed for dealing with the large quantities of data produced by Twitter, Facebook, and blogs.

The platform was conceptualised at the start of this study, and then built into a usable and functioning social media analytics environment. A paper[57] on the SocialSTORM platform was presented at WORLDCOMP'12, and was also covered The Financial Times[58]. The SocialSTORM platform was used in the study for preliminary investigations, allowing for the identification of how to most-efficiently access; process; and store the data produced by Twitter. It also produced early results which guided the selection of the analytics methodologies used and the study's dependency parameters. Furthermore, the knowledge gained from this trial and error phase was used as input to structure and build a dedicated Twitter collection system (denoted: 'Twitter Collection Framework' or TCF), used for data-collection throughout the study. This platform recorded from both Twitter's Sample[b] and Filter[c] data streams, and was used to produce the raw data used in exploring both of the hypotheses in this investigation. Both platforms had access to Twitter's 10% 'Gardenhose'[d] data-feed, which can produce up to 40 million messages per day.

[a] http://social2.cs.ucl.ac.uk:8080/

[b] Twitter's Sample feed contains a random sample of all Tweets sent globally.

[c] Twitter's Filter feed allows programmatic filtering of the Tweets, prior to their return to the user. For example, the feed can be configured to only return Tweets which contain a particular string.

[d] 'Gardenhose' is the name Twitter gives to its publicly-available increased-access data-feed that can be used for academic research. It contains 10% of all of Twitter's messages streamed through its API

4.1 SocialSTORM, UCL's social media analytics engine

The SocialSTORM platform was built as part of the study, but was a departmental-wide initiative to gain access to social media data sources for research purposes. It also acted as a test-bed that guided the development of the Twitter Collection Framework (TCF) which was used to access Twitter data for the study, and as a source of data for preliminary results.

As discussed earlier, SocialSTORM is a cloud-based platform which facilitated the acquisition of text-based data from online sources such as Twitter, Facebook, respected blogs, RSS media and traditional news-sources. As a central-hub for social media analytics, the system included facilities to upload and run Java-coded simulation models to analyse the previously-collected data. SocialSTORM also had connectivity to UCL's ATRADE[a] platform which provided further quantitative finance and economic data.

The platform consisted of infrastructure and tools to facilitate data acquisition, database connectivity, and varying levels of access and administration along with data repositories for long and short-term data storage. The platform was able to operate both in an 'historical' mode which utilised data already stored at the time of running the desired simulation, and a 'live' mode which operated on a near real-time stream[b] of data which was continually monitored throughout the simulation. These differing modes permitted the evaluation of models in an accelerated manner[c] when executed on historic data, before being evaluated at real-time speeds when processing live data.

The SocialSTORM platform resided on a leased server from UCL's Computer Science department, but was designed for migration to a cloud computing environment[d]. This environment consisted of 16 nodes each with the following specification: 15,000rpm 600GB hard drive, 32GB RAM and one 3.2GHz quad-core Intel Xeon e3-1200 processor. The nodes were interlinked by 10GbE (10 Gigabit Ethernet) connections and the entire system was backed-up daily onto tape storage for up to 3 months. SocialSTORM's storage capacity, when migrated to this cluster was 8TB with 512GB

selected at random, but access to this feed needs to be arranged with Twitter on a user-by-user basis. The standard feed, which requires no prior arrangement with Twitter to access, offers up to a 1% random sample of all of Twitter's messages.

[a] A cloud-based computational finance environment built by UCL that supports real and virtual trading; with terabytes of financial data to support research into algorithmic trading and risk. http://vtp.cs.ucl.ac.uk/atrade

[b] < 1 second latency from the time of Tweet creation to the time of Tweet storage.

[c] Quicker than real-time

[d] Cloud computing involves distributing computational processes and/or storage across multiple systems networked together to share, and make use of, the computing power provided by multiple machines.

of available RAM. SocialSTORM was fully scalable – additional nodes could be added to increase system storage and performance on an as-needed basis. The funding for this hardware was provided by UCL's Computer Science Department.

This particular hardware setup was chosen for the purposes of integrating SocialSTORM with Apache's Hadoop[a], a software library and framework that allows for the distributed processing of large data sets.

SocialSTORM inherited its architectural design from UCL's ATRADE system, and so allowed for easy integration between the two systems. The following is an outline of the key components of the SocialSTORM system.

Connectivity Engines – Various connectivity modules communicated with the external data sources, including Twitter & Facebook's APIs, financial blogs and various RSS news feeds. Data were fed into SocialSTORM in real-time and included a 10% random sample of all public updates from Twitter, as well as filtered data streams selected from a rich dictionary of stock symbols, currencies and other economic keywords; providing gigabytes of text-based data every day.

Messaging Bus – This served as the internal communication layer which accepted the incoming data streams (messages) from the various connectivity engines, parsed these and wrote the various data to the appropriate tables of the main database.

Data Warehouse – This was a MySQL[b] relational database[c], chosen for its open-source[d] nature and its particular ability to ingest high-volume, high-velocity data. It housed terabytes of text-based entries which were accompanied by all associated metadata. Entries were organised by source and accurately time-stamped with the time of publication, as well as being tagged with topics for easy retrieval by simulation models.

Simulation Manager – This terminal provided the external API for end-users to interact with the data for the purposes of analysis, including a web-based GUI via which users could upload a Java-coded simulation model to perform the desired analysis on

[a] http://www.hadoop.apache.org

[b] MySQL is a cross-platform relational database management system owned by Oracle Corporation. It is famed for its simplicity, inexpensiveness, scalability and speed.

[c] A relational database is a collective set of datasets organised by tables which have a defined relationship between each other, permitting scalability, speed and the efficient use of storage space.

[d] In software design, open-source denotes a development model which provides universal access via free license for use as well as for access to the program's code.

the data. The Simulation Manager facilitated all client-access to the data warehouse, and also allowed users to upload their own datasets for simultaneous analysis alongside the social media data.

In summary, the SocialSTORM platform included acquisition and access to terabytes of social media data from a variety of sources, as well as a cloud-based simulation environment for historical and real-time monitoring of global news and opinions taken from the world’s most popular social networking and news sites.

SocialSTORM queried and monitored its data-acquisition APIs in real-time, reading updates as they were streamed and wrote these directly to its database. The latency between a message being published to Twitter (as an example) and subsequently being stored in the database was less than 1 second; even when using batch inserts to increase efficiency. Typically, the system wrote 4,000 entries to the database every second.

From Twitter, the system retrieved up to 40 million messages per day as a ‘random sample’ of all public updates, plus up to 2 million messages daily containing hundreds of specific financial and economic keywords selected by the platform’s development team. From Facebook, a proprietary method of retrieving a random sample of all public updates was used which returns up to 2 million updates per day. The SocialSTORM system also had programmatic scrapers for 15 finance-related blogs, as well as a number of official news services which, together contributed over 1,000 daily entries to the database. The data sources resulted in the collection of approximately 5GB of data per day.

User-privacy was taken seriously by the platform’s development team. Although the data retrieved from the web is in the public domain, it remains property of the data provider and is therefore not redistributable in accordance with content license agreements. To enable analysis of social media data by third parties, SocialSTORM was designed as a black-box research environment from which the raw data cannot be downloaded. Instead, the system was accessible via a graphical web interface, which is shown in Figure 4.

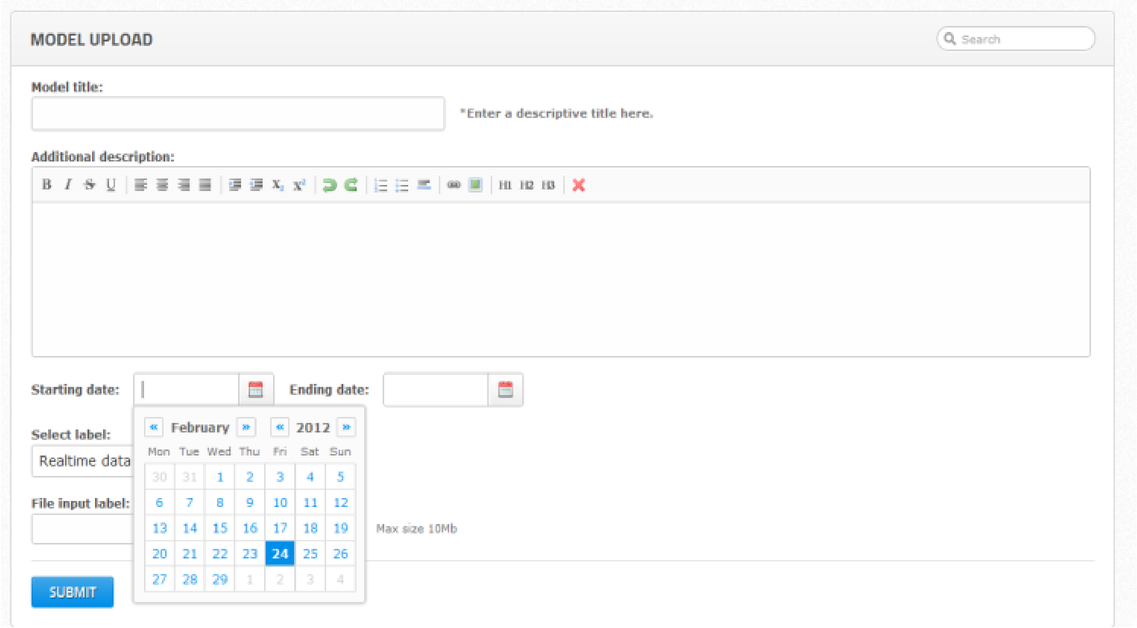


FIGURE 4: MODEL-UPLOAD FORM FOR SOCIALSTORM SIMULATIONS

Figure 4 shows the graphical web-based user interface via which subscribed users were able to upload their own java-coded simulation models which would guide the analysis of the data stored by SocialSTORM. The significance of the creation of the graphical user interface is the ability for SocialSTORM's users to monitor the progress of their simulation models from any web-connected device with a browser without having to stop the models and inspect the results in third-party software to identify incorrect functionality of the models.

Models were uploaded to SocialSTORM as .jar files, which also included any packages on which the code was dependent. The simulation environment then looked for a particular method, similar to `Main()`, which defined the appropriate parameters via which to interface with SocialSTORM's datasets. Instructions on how to ensure that models were compliant with the platform are detailed in the SocialSTORM user manual, which is available in the Appendix (see Chapter 11.1).

Once a simulation was complete, users would visualise results using SocialSTORM's output GUI (an example of which is shown in Figure 5); export the results to Microsoft Excel; or use an output API to retrieve the results programmatically for further analysis. Data exported to Microsoft Excel could be linked to automatically update in a spread-sheet's cells.

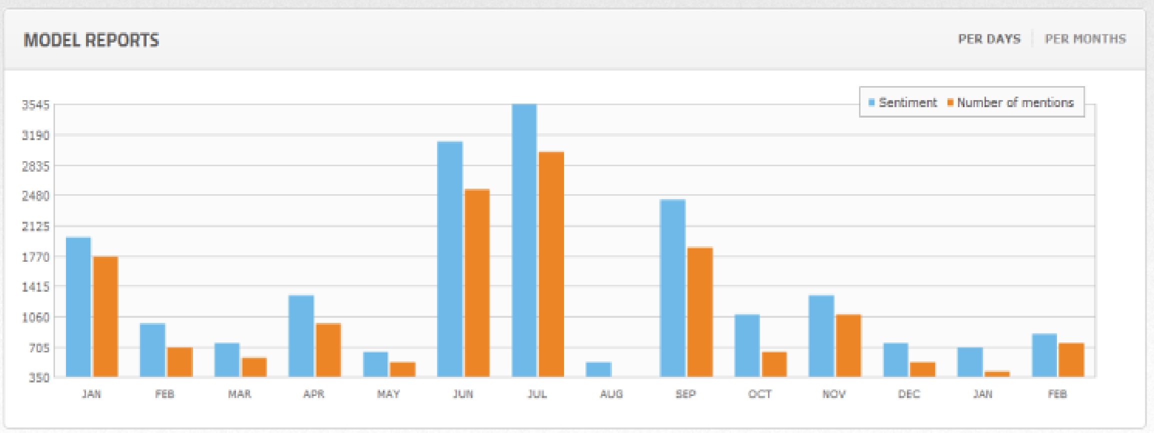


FIGURE 5: EXAMPLE OF RESULTS VISUALISED IN SOCIALSTORM'S WEB INTERFACE

The creation of a system for visualising models' output in real-time via a web-based system, as seen in Figure 5, permitted the end-user to easily see a simulation's output, and therefore assess its performance without having to first export the data to a separate software package for visualisation. This functionality was built to improve the real-world usability of the system by decreasing dependency on third-party software for data-visualisation.

The development of SocialSTORM was used in this study as a guide for the construction of a dedicated software package suited specifically for the filtering of Twitter data with regards to the study's hypotheses (see Chapter 4.2), and as a source of data for preliminary results (as discussed in Chapters 5.3 and 5.6.1).

4.2 Twitter Collection Framework (TCF)

To explore the two hypotheses set out by this study (see Chapter 1.3), a programmatic method for collecting Twitter data was needed which met the following criteria:

1. The ability to filter Twitter's data by geographical and/or string criteria according to the Twitter filters as listed in Table 1;
2. The ability to receive the filtered data produced by Twitter's network at a speed sufficient to prevent the formation of backlogs of undelivered data, which would result in disconnection from Twitter's network;

3. The ability to store the filtered data in such a manner that invisible new-line characters[a] within the Tweets did not result in the writing of the data to incorrect locations within the output files;

4. The ability to reliably connect, maintain a connection of sufficient bandwidth, and disconnect from Twitter's network according to the network's strict protocols;

5. The ability to store the filtered data in a format that could be subsequently read by a mathematical analytics environment.

A platform, known as the Twitter Collection Framework (TCF) was written to address the above requirements. Created in Java and residing within the Eclipse Integrated Development Environment[b], the program was written to simultaneously connect to Twitter's Sample API and to Twitter's Filter API using a multithreaded approach. This permits the user to record both all of the data pushed through Twitter's Gardenhose Feed, and data which matches a particular set of filters. These filters can either be keywords, e.g., "iPhone", Twitter user IDs e.g., "@BritishMonarchy" or pairs of longitude and latitude coordinates which bound a geographical area from which Tweets should be recorded. For example, the coordinates '40,-74' and '41,-73' represent the South-West and North-East coordinates which bound New York City.

The TCF resided within a single dedicated server and a full copy of the code underpinning the framework is available in the Appendix (see Chapter 11.1). The TCF's functionality was controlled by an XML[c] file, which contained a list of string, and/or geographical-location filters ('Twitter filters') which define the criteria by which the TCF filters Twitter's incoming data streams. By using this XML control file, the TCF was able to filter incoming Tweets based on the locations they are sent from, and/or string combinations in accordance to Table 1. String combinations can be in the form of:

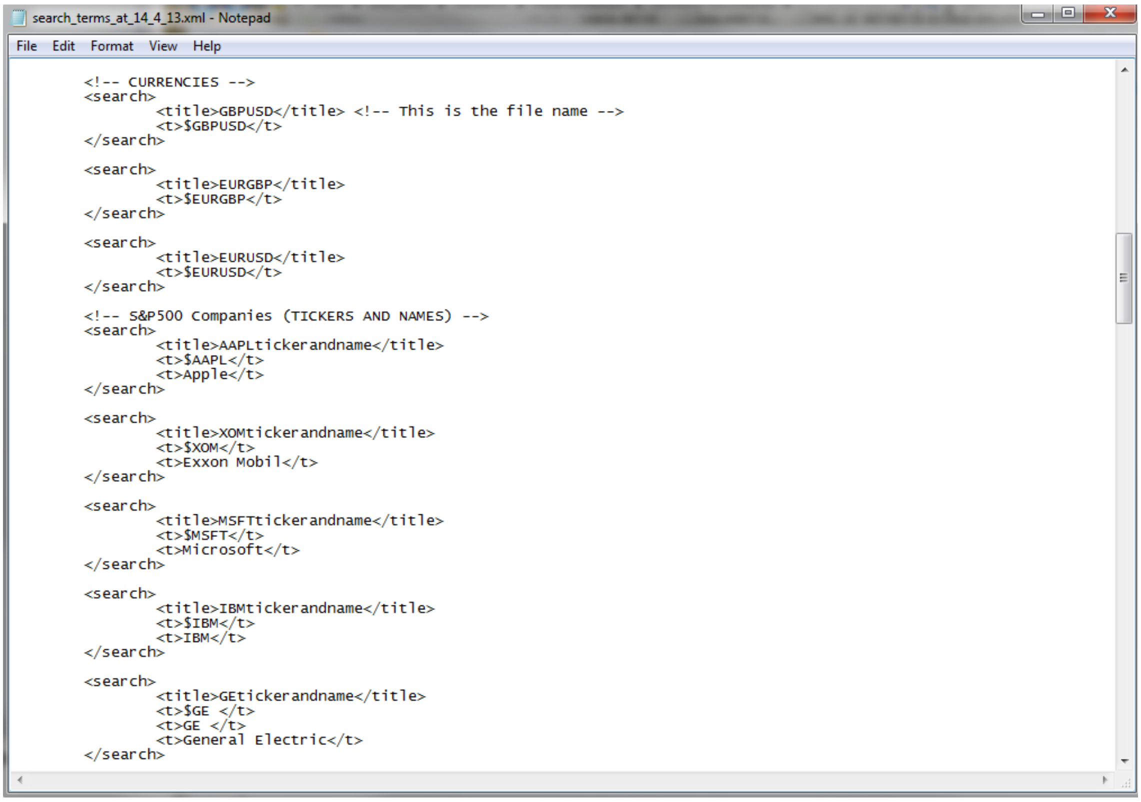
[a] During the construction of SocialSTORM, it was determined that Tweets frequently contain invisible characters which cause the creation of new lines in the text strings. This is the result of the use of 3rd party programs to contribute to Twitter's network. Such characters cannot be seen visually, but when processed programmatically, they caused misalignment resulting in the storage of Tweets to incorrect locations within the output files.

[b] A multi-language development environment http://www.eclipse.org/

[c] Extensible Markup Language – a system for encoding documents that are readable by both humans and machines.

- AND statements, e.g., “$AAPL” AND “apple”.
- OR statements, e.g., “work” OR “play”.
- Combinations:
    - [“$AAPL” AND “apple”] OR [“work” OR “play”].

An example of the XML filter file can be seen below:


search_terms_at_14_4_13.xml - Notepad
File Edit Format View Help

```
        <!-- CURRENCIES -->
        <search>
                <title>GBPUSD</title> <!-- This is the file name -->
                <t>$GBPUSD</t>
        </search>

        <search>
                <title>EURGBP</title>
                <t>$EURGBP</t>
        </search>

        <search>
                <title>EURUSD</title>
                <t>$EURUSD</t>
        </search>

        <!-- S&P500 Companies (TICKERS AND NAMES) -->
        <search>
                <title>AAPLtickerandname</title>
                <t>$AAPL</t>
                <t>Apple</t>
        </search>

        <search>
                <title>XOMtickerandname</title>
                <t>$XOM</t>
                <t>Exxon Mobil</t>
        </search>

        <search>
                <title>MSFTtickerandname</title>
                <t>$MSFT</t>
                <t>Microsoft</t>
        </search>

        <search>
                <title>IBMtickerandname</title>
                <t>$IBM</t>
                <t>IBM</t>
        </search>

        <search>
                <title>GEtickerandname</title>
                <t>$GE </t>
                <t>GE </t>
                <t>General Electric</t>
        </search>
```

FIGURE 6: EXAMPLE OF THE XML-BASED TWITTER-FILTER CONTROL FILE

In Figure 6 above, the text contained within `<search>…</search>` tags represents a particular Twitter Filter, providing a straightforward system for managing and operating the TCF. Each `<title>…</title>` tag contains the filename of the resultant .txt file corresponding to a single twitter Filter. Each `<t>…</t>` tag contained a string to be matched. These features addressed the first of the aforementioned criteria.

The results produced by each Twitter filter were stored in a separate .txt file. There is no internal software limit to how many search filters can be used. A hardware limit does exist though: as the number of filters increases, the TCF’s requirement for system memory also increases. By trial and error, it was determined that up-to 50 filters can be sustained indefinitely on a machine with 4GB of RAM. Disconnections due to memory-allocations became an issue as the number of Twitter filters increased above 55. Therefore, with a contingency of 5 Twitter filters, the TCF's maximum capacity was designated to be 50. This capacity sufficiently addressed the second of the aforementioned criteria.

The TCF was developed to store raw Tweets filtered based on a set of geographical and/or string filters, the sentiments and volumes of which would be analysed in the future. It was discovered during the development of SocialSTORM that the volume of data produced by programmatically accessing Twitter was too large for a non-distributed system to handle indefinitely, causing routine memory-related outages for both of the TCF and SocialSTORM, thus not meeting the second of the aforementioned criteria. Unfiltered, Twitter's diminished (10%) Gardenhose Feed yielded up to 40 million messages per day. In the case of both the TCF and SocialSTORM, the systems could not support 50 Twitter filters indefinitely due to memory issues if the underlying raw strings were stored. To cope with this volume of data without being subject to memory-related issues, the TCF had to be written such that sentiment classification took place at the point of Tweet collection. The exact nature of the sentiment classification methodologies used in this research project is discussed in Chapter 5.1. Furthermore, the implementation of this decision addressed the third of the aforementioned criteria.

For validation, the underlying (discarded) Tweets used in this study can be obtained from the databases of commercial-grade platforms such as Topsy[a], which offers for-pay access to historic Tweets.

To address the fourth of the aforementioned criteria, the TCF system required an internet connection that would support the uninterrupted delivery of all Tweets in real time[b], otherwise a backlog occurs and not all Tweets are delivered on time. Twitter's API documentation strongly advises against such situations since a connected system's sustained inability to receive all the Tweets being fed through will result in a forced disconnection. Such disconnections would require the system to automatically reconnect via a strict protocols[c] to avoid barring by Twitter's network. The study used

[a] Topsy, available at http://www.topsy.com, is a social search and analytics company founded in January 2007 which designed a system for accessing all historic Tweets. This cloud-computing system was designed and developed over five years after securing funding of $35.2 million, before being sold to Apple, Inc. for $200 million in 2003.

[b] It has been found via testing that a 10mBit/s connection is sufficient.

[c] https://dev.twitter.com/docs/auth

Twitter4J[a], an open-source Java library designed specifically to follow these protocols, in order to deal with connections, disconnections and reconnections to Twitter's APIs in accordance with Twitter's guidelines. By implementing this library in the TCF, a reliable connection to Twitter's APIs would be guaranteed thus minimising the accidental and unnecessary omission of any Twitter data in the study's analyses. In such a manner, the TCF met the fourth of the aforementioned criteria.

Finally, the TCF was designed to store the Tweet outputs yielded by these term-matching and location-matching filters to .txt files, which could then be easily read and analysed in a mathematical analytics environment, e.g., MATLAB[b] – thus meeting the fifth of the aforementioned criteria.

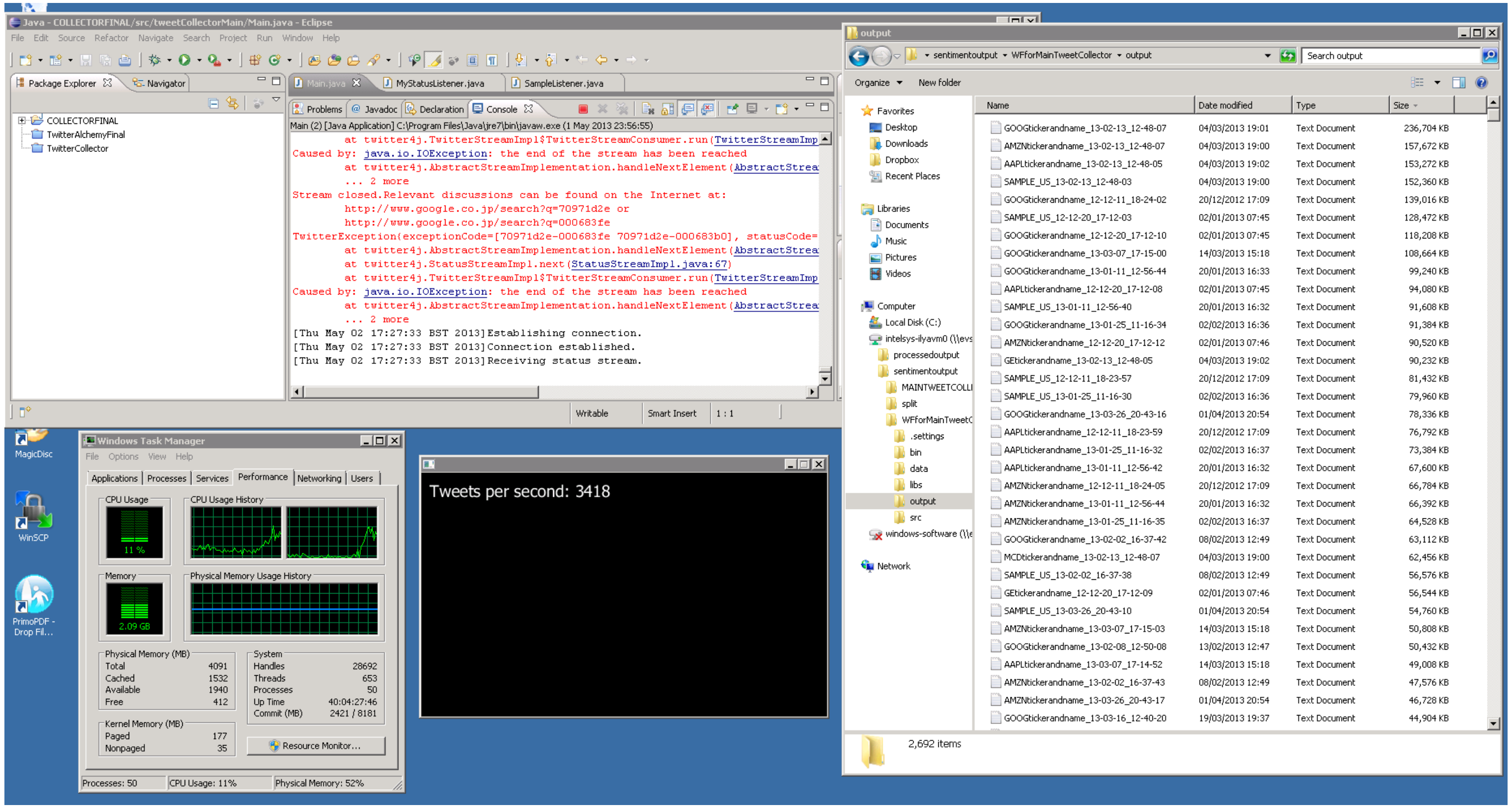
Provided that the internet connection powering the TCF is of sufficient bandwidth, the platform can comfortably deal with Twitter's 10% Gardenhose Feed. Its maximum collection and sentiment classification performance has been benchmarked at 13,315 Tweets per second. In comparison, Twitter's 10% Gardenhose Feed produced an average of 578 Tweets per second in 2013 (calculated from the delivery of an average of 500 million messages per day from Twitter's 100% Firehose Feed[59]). However, it should be noted that in exceptional circumstances, the 10% Gardenhose Feed has been known to deliver message rates an order of magnitude greater for periods of up to a few seconds[c]. The TCF could therefore reliably deal with the 10% Gardenhose Feed and any exceptional message-volume rates, but would only sustain an average daily volume of messages passed through Twitter's 100% Firehose during non-exceptional periods – the volume of messages during exceptional high-activity periods could be too great for the TCF to sustain.

A screenshot of the TCF system is provided in Figure 7:

[a] http://twitter4j.org

[b] A numerical computing environment suitable for processing and analysing large numerical or textual data-sets.

[c] For example, the delivery of news of unprecedented importance via Twitter, e.g., the Boston Marathon Bombings of 15th April 2013 which resulted in a peak of 4,000 Tweets per second.



```
Main (2) [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (1 May 2013 23:56:55)
        at twitter4j.TwitterStreamImpl$TwitterStreamConsumer.run(TwitterStreamImp
Caused by: java.io.IOException: the end of the stream has been reached
        at twitter4j.AbstractStreamImplementation.handleNextElement(AbstractStrea
        ... 2 more
Stream closed.Relevant discussions can be found on the Internet at:
        http://www.google.co.jp/search?q=70971d2e or
        http://www.google.co.jp/search?q=000683fe
TwitterException{exceptionCode=[70971d2e-000683fe 70971d2e-000683b0], statusCode=
        at twitter4j.AbstractStreamImplementation.handleNextElement(AbstractStrea
        at twitter4j.StatusStreamImpl.next(StatusStreamImpl.java:67)
        at twitter4j.TwitterStreamImpl$TwitterStreamConsumer.run(TwitterStreamImp
Caused by: java.io.IOException: the end of the stream has been reached
        at twitter4j.AbstractStreamImplementation.handleNextElement(AbstractStrea
        ... 2 more
[Thu May 02 17:27:33 BST 2013]Establishing connection.
[Thu May 02 17:27:33 BST 2013]Connection established.
[Thu May 02 17:27:33 BST 2013]Receiving status stream.
```

Physical Memory (MB)	
Total	4091
Cached	1532
Available	1940
Free	412
Kernel Memory (MB)	
Paged	177
Nonpaged	35

System	
Handles	28692
Threads	653
Processes	50
Up Time	40:04:27:46
Commit (MB)	2421 / 8181

Processes: 50 CPU Usage: 11% Physical Memory: 52%

Tweets per second: 3418

Name	Date modified	Type	Size
GOOGtickerandname_13-02-13_12-48-07	04/03/2013 19:01	Text Document	236,704 KB
AMZNtickerandname_13-02-13_12-48-07	04/03/2013 19:00	Text Document	157,672 KB
AAPLtickerandname_13-02-13_12-48-05	04/03/2013 19:02	Text Document	153,272 KB
SAMPLE_US_13-02-13_12-48-03	04/03/2013 19:00	Text Document	152,360 KB
GOOGtickerandname_12-12-11_18-24-02	20/12/2012 17:09	Text Document	139,016 KB
SAMPLE_US_12-12-20_17-12-03	02/01/2013 07:45	Text Document	128,472 KB
GOOGtickerandname_12-12-20_17-12-10	02/01/2013 07:45	Text Document	118,208 KB
GOOGtickerandname_13-03-07_17-15-00	14/03/2013 15:18	Text Document	108,664 KB
GOOGtickerandname_13-01-11_12-56-44	20/01/2013 16:33	Text Document	99,240 KB
AAPLtickerandname_12-12-20_17-12-08	02/01/2013 07:45	Text Document	94,080 KB
SAMPLE_US_13-01-11_12-56-40	20/01/2013 16:32	Text Document	91,608 KB
GOOGtickerandname_13-01-25_11-16-34	02/02/2013 16:36	Text Document	91,384 KB
AMZNtickerandname_12-12-20_17-12-12	02/01/2013 07:46	Text Document	90,520 KB
GEtickerandname_13-02-13_12-48-05	04/03/2013 19:02	Text Document	90,232 KB
SAMPLE_US_12-12-11_18-23-57	20/12/2012 17:09	Text Document	81,432 KB
SAMPLE_US_13-01-25_11-16-30	02/02/2013 16:36	Text Document	79,960 KB
GOOGtickerandname_13-03-26_20-43-16	01/04/2013 20:54	Text Document	78,336 KB
AAPLtickerandname_12-12-11_18-23-59	20/12/2012 17:09	Text Document	76,792 KB
AAPLtickerandname_13-01-25_11-16-32	02/02/2013 16:37	Text Document	73,384 KB
AAPLtickerandname_13-01-11_12-56-42	20/01/2013 16:32	Text Document	67,600 KB
AMZNtickerandname_12-12-11_18-24-05	20/12/2012 17:09	Text Document	66,784 KB
AMZNtickerandname_13-01-11_12-56-44	20/01/2013 16:32	Text Document	66,392 KB
AMZNtickerandname_13-01-25_11-16-35	02/02/2013 16:37	Text Document	64,528 KB
GOOGtickerandname_13-02-02_16-37-42	08/02/2013 12:49	Text Document	63,112 KB
MCDtickerandname_13-02-13_12-48-07	04/03/2013 19:00	Text Document	62,456 KB
SAMPLE_US_13-02-02_16-37-38	08/02/2013 12:49	Text Document	56,576 KB
GEtickerandname_12-12-20_17-12-09	02/01/2013 07:46	Text Document	56,544 KB
SAMPLE_US_13-03-26_20-43-10	01/04/2013 20:54	Text Document	54,760 KB
AMZNtickerandname_13-03-07_17-15-03	14/03/2013 15:18	Text Document	50,808 KB
GOOGtickerandname_13-02-08_12-50-08	13/02/2013 12:47	Text Document	50,432 KB
AAPLtickerandname_13-03-07_17-14-52	14/03/2013 15:18	Text Document	49,008 KB
AMZNtickerandname_13-02-02_16-37-43	08/02/2013 12:49	Text Document	47,576 KB
AMZNtickerandname_13-03-26_20-43-17	01/04/2013 20:54	Text Document	46,728 KB
GOOGtickerandname_13-03-16_12-40-20	19/03/2013 19:37	Text Document	44,904 KB

FIGURE 7: SCREENSHOT OF THE TWITTER COLLECTION FRAMEWORK (TCF) IN OPERATION

Figure 7 shows examples of: disconnection-related error-handling procedures (top-left); the resultant .txt storage files which contain sentiments of the Tweets filtered in by the

TCF; the system's CPU usage; and an indication of the number of Tweets processed per second.

4.3 Relationships explored in the study based on the evaluation of Twitter data

Forty-four financial-instrument/Twitter-Filter combinations were set-up to collect social media data using the TCF. With regards to the Hypothesis One (see Chapter 1.3), these financial-instrument/Twitter-Filter combinations were used to answer the following questions:

1. What is the relationship between a random sample of Tweets deemed to be from the US, and the returns of a leading US stock index, specifically the S&P500?
2. What is the relationship between a random sample Tweets deemed to be from the UK, and returns of a leading UK stock index, specifically the FTSE100?

To evaluate these relationships, the sentiments and message volumes of string-unfiltered Tweets from the UK were evaluated against returns of FTSE100 Futures[a] and CFDs[b]. Similarly, the sentiments and message volumes of string-unfiltered Tweets from the US were evaluated against returns of S&P500 Futures and CFDs. Price data for Futures were obtained from Fulcrum Asset Management[c]. Price data for CFDs were obtained from the Swiss foreign-exchange bank and marketplace Dukascopy[d].

With regards to Hypothesis Two (see Chapter 1.3), the aforementioned financial-instrument/Twitter-Filter combinations were used to answer the following questions:

1. What are the relationships between the sentiments and message volumes of Tweets filtered by the following currency pairs, and their returns: GBPUSD and EURGBP?

[a] This is a market-tradable financial contract between two parties to buy or sell a specified asset at a price agreed upon today, but with delivery and payment occurring in the future.

[b] CFD (or Contract For Difference) is a market-tradable financial contract between two parties to exchange the difference between the current value of a specified asset and its value at a contract time. It is used to speculate about the underlying asset's price movement without the need to own the asset itself.

[c] http://www.fulcrumasset.com/

[d] http://www.dukascopy.com/

2. What are the relationships between the sentiments and message volumes of Tweets filtered by the top constituents of the S&P500, and the top constituents of the Dow Jones Industrial Average, and their returns?

To evaluate the relationships of the currency pair exchange rates with Twitter data, two string-based Twitter filters were set-up to monitor for mentions of “$GBPUSD” and “$EURUSD”[a]. The sentiments and volumes of these messages were evaluated against the returns of their respective currency pairs, both for Futures prices and CFD prices (provided by Fulcrum Asset Management and Dukascopy, respectively).

To evaluate the relationships of the returns of the top constituents of the S&P500 and the Dow Jones Industrial Average, a series of string-based Twitter filters were set-up to filter Tweets mentioning the aforementioned assets’ industry Ticker-IDs, as according to Table 1. Another set of Twitter filters was set-up to filter Tweets mentioning these assets’ industry Ticker-IDs AND/OR their Company Names. In each case, the sentiments and message volumes of the data produces by these Twitter filters were evaluated against the returns the respective securities’ CFDs.

For certain companies, it was necessary to include alternative-spellings for the companies’ names. For example, for the company Coca-Cola, Co., a Twitter filter was used to filter in all Tweets which matched either “Coca-Cola” OR “Coca Cola” since both versions are used to refer to this company. This was made especially necessary since Twitter’s search API does not support wildcard matching[b].

[a] Here, the “$” symbol is used in the financial industry to denote a particular tradable asset via the use of its industry ticker. For example, the industry ticker for the shares of Apple, Inc. is “$AAPL”.

[b] A wildcard (“*”) is a character which can be used as a substitute for any of a defined subset of possible characters. For example, the string “heat**” could match the words “heated” and “heater”, if wildcard filtering were permitted on Twitter’s network.

Filter ID	Instrument	Filter type	Filter
1	Apple, Inc. CFDs	Ticker-ID AND/OR Company Name	$AAPL AND/OR “Apple”
2	Apple, Inc. CFDs	Ticker-ID	$AAPL
3	Amazon.com, Inc. CFDs	Ticker-ID AND/OR Company Name	$AMZN AND/OR “Amazon”
4	Amazon.com, Inc. CFDs	Ticker-ID	$AMZN
5	American Express, Co. CFDs	Ticker-ID AND/OR Company Name	$AXP AND/OR “American Express”
6	Bank of America, Corp. CFDs	Ticker-ID AND/OR Company Name	$BAC AND/OR “Bank of America”
7	Bank of America, Corp. CFDs	Ticker-ID	$BAC
8	Cisco Systems, Inc. CFDs	Ticker-ID AND/OR Company Name	$CSCO AND/OR “Cisco”
9	EURUSD CFDs	Ticker-ID	$EURUSD
10	EURUSD Futures	Ticker-ID	$EURUSD
11	GBPUSD CFDs	Ticker-ID	$GBPUSD
12	GBPUSD Futures	Ticker-ID	$GBPUSD
13	General Electric, Co. CFDs	Ticker-ID AND/OR Company Name	$GE AND/OR “GE” AND/OR “General Electric”
14	General Electric, Co. CFDs	Ticker-ID	$GE
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	$GOOG AND/OR “Google”
16	Google, Inc. CFDs	Ticker-ID	$GOOG
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	$HD AND/OR “Home Depot”
18	Hewlett Packard, Co. CFDs	Ticker-ID AND/OR Company Name	$HPQ AND/OR “Hewlett-Packard” AND/OR “Hewlett Packard”
19	Hewlett Packard, Co. CFDs	Ticker-ID	$HPQ
20	IBM, Corp. CFDs	Ticker-ID AND/OR Company Name	$IBM AND/OR “IBM”
21	IBM, Corp. CFDs	Ticker-ID	$IBM
22	Intel, Corp. CFDs	Ticker-ID AND/OR Company Name	$INTC AND/OR “Intel”
23	Intel, Corp. CFDs	Ticker-ID	$INTC
24	Johnson & Johnson, Co. CFDs	Ticker-ID AND/OR Company Name	$JNJ AND/OR “Johnson & Johnson” AND/OR “Johnson and Johnson”
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	$JPM AND/OR “JPMorgan” AND/OR “JP Morgan”
26	J.P. Morgan, Inc. CFDs	Ticker-ID	$JPM
27	Coca-Cola, Co. CFDs	Ticker-ID AND/OR Company Name	$KO AND/OR “Coca-Cola” AND/OR “Coca Cola”
28	Coca-Cola, Co. CFDs	Ticker-ID	$KO
29	McDonald’s, Corp. CFDs	Ticker-ID AND/OR Company Name	$MCD AND/OR “McDonald’s” AND/OR “McDonalds”
30	McDonald’s, Corp. CFDs	Ticker-ID	$MCD
31	3M, Co. CFDs	Ticker-ID AND/OR Company Name	$MMM AND/OR “3M”
32	Microsoft, Corp. CFDs	Ticker-ID AND/OR Company Name	$MSFT AND/OR “Microsoft”
33	Microsoft, Corp. CFDs	Ticker-ID	$MSFT
34	Oracle, Corp. CFDs	Ticker-ID & Company Name	$ORCL AND/OR “Oracle”
35	Oracle, Corp. CFDs	Ticker-ID	$ORCL
36	FTSE100 Index CFDs	UK Geo via coordinate-matching	String-unfiltered UK Tweets
37	FTSE100 Index Futures	UK Geo via coordinate-matching	String-unfiltered UK Tweets
38	S&P500 Index CFDs	US Geo via coordinate-matching	String-unfiltered US Tweets
39	S&P500 Index Futures	US Geo via coordinate-matching	String-unfiltered US Tweets
40	AT&T, Inc. CFDs	Ticker-ID AND/OR Company Name	$T AND/OR “AT&T”
41	AT&T, Inc. CFDs	Ticker-ID	$T
42	Wal-Mart, Inc. CFDs	Ticker-ID AND/OR Company Name	$WMT AND/OR “Wal-Mart” AND/OR “Wal Mart”
43	Exxon Mobil, Corp. CFDs	Ticker-ID AND/OR Company Name	$XOM AND/OR “Exxon Mobil”
44	Exxon Mobil, Corp. CFDs	Ticker-ID	$XOM

TABLE 1: LIST OF TWITTER FILTERS USED IN THE STUDY

5 ANALYSING TWITTER DATA

This chapter details the mathematical methodologies used to evaluate the extent to which Twitter data can lead financial data.

5.1 Sentiment analysis

As discussed in the literature review (Chapter 3), previous work on the analysis of Tweets for correlation with financial market events has shown that a statistically-significant relationship can exist between Tweet volumes on particular equities, and their market trading volumes[11]. Such correlations are an indication of the existence of a relationship linking the two data sets together, but this sort of analysis does not attempt to address what correlation the meaning of the underlying messages can have with financial market data. As discussed in Chapter 3.3, the mood of a Tweet's contents can be quantified in a programmatic sense by using sentiment analysis – and in such a manner an automated quantitative analysis of the meaning of Tweets can be performed.

A sentiment analysis classifier is therefore required for the study. The necessary criteria for the tool are:

- Accuracy. The classifier must be designed to accurately rank the nature of the text present in social media data, i.e., text which does not necessarily subscribe to the correct spelling and grammatical rules of the English language;
- Convenience and speed of use in mass data analysis. Specifically, the sentiment classifier must be accessible programmatically rather than via manual procedures only. This is necessary for practical applicability of the study's findings given the volume of social media data involved: this study will not have meaning unless the classifier employed can be used in a manner that is near-enough to real-time to allow for practical trading implementation – a consideration that is further addressed in Chapters 6.1 and 6.3.3;
- Transparency of internal operations given the academic intentions of the study. The sentiment classifier's principles of operation must be visible, upon request, during the classification of any text string.

This study is centred on the implementation of a sentiment analysis tool which offered the greatest known accuracy in ranking the text types fed through Twitter's network.

Four tools were evaluated for this purpose, and their characteristics were further compared to other techniques, via extant published results in literature. The tools examined were:

AlchemyAPI[a]: This is a commercial language analysis system which offers access to Natural Language Processing tools, of which one is a sentiment classifier. With reference to the aforementioned criteria:

- AlchemyAPI's sentiment classifier is not suited to the nature of text used in social media. It has been shown to ignore misspellings of common words rather than classify them[60] – an issue that cannot be overlooked due to intended misspellings of text on Twitter caused by the message-length constraints of the network[61];
- AlchemyAPI's sentiment classifier is programmatically accessible, however it is a commercial system, and therfore incurs a per-string classification charge;
- AlchemyAPI's sentiment classifier is non-transparent, meaning that the methodology behind determining the sentiment of any batch of text cannot be seen by the user.

Some initial investigation of the system's capabilities was performed via its integration with the TCF. However, due to the fact that its internal operations are not transparent, it was deemed unsuitable for replication of results in the present academic study. Furthermore, its unsuitability to ranking informal English makes the system unsuited to Twitter vernacular.

Custom Naïve Bayes Classifier[62]: As part of the SocialSTORM project, a sentiment analysis classifier was built into the software system by an MSc student (Long, M. *Sentiment analysis using a Naïve Bayes classifier*. MSc Thesis as part of the MSc in Computer Science at University College London, 2012). The classifier was designed to mimic the output produced by AlchemyAPI. With reference to the aforementioned criteria:

[a] http://www.alchemyapi.com/

- Since the Custom Naïve Bayes Classifier was designed to mimic AlchemyAPI's sentiment classifier, which has been shown[60] to not accurately rank informal text, it cannot be suited to the nature of social media vernacular. Furthermore, the Custom Naïve Bayes Classifier was only up to 35% accurate at mimicking AlchemyAPI's sentiment classifier[62];

- The Custom Naïve Bayes Classifier was programmatically accessible.

LIWC (Linguistic Inquiry and Word Count)[a]: This is a research-centric dictionary-term matching sentiment classifier which uses a 9906-word corpus to calculate the prevalence of emotion in text according to a range of emotions/personal concerns[b], and linguistic processes[c]. With reference to the aforementioned criteria:

- This package's strength is its accuracy in ranking formal English since the corpus of words contained within, and their attributions to various emotions, is the result of substantial research in the field of Natural Language Processing which has been scrutinised by panels of human judges[51]. However, the system does not contain the capability for ranking informal text. For example: its corpus does not contain the common misspellings of common English words; and the platform does not take into account the negating effects of negators (e.g., "not bad"). It is instead geared towards ranking the prevalence of emotion in longer prose[51];

- The LIWC package is slow from a usability point of view as it cannot be accessed programmatically, instead relying on the user to feed in text manually into a graphical user interface;

- The LIWC package's classification methodology is fully transparent, meaning that the methodology behind determining the sentiment of any batch of text can be seen by the user.

[a] http://www.liwc.net/

[b] Social processes (family/friends/humans). Affective processes (positive emotion, negative emotion, anxiety, sadness, fear). Cognitive processes (insight, causation, discrepancy). Perceptual processes (seeing, hearing, feeling). Biological processes (body, health, sexuality). Relativity (motion, space, time). Personal concerns (work, leisure, home, money, religion, death).

[c] Pronouns, articles, verbs. Tense identification. Adverbs, prepositions, conjunctions, quantifiers, profanities.

SentiStrength[a]: This is a research-oriented sentiment classifier designed for ranking short informal text in the English language. The system has seen previous implementation in ranking the sentiments of Tweets in academic[63] and commercial exercises[b64,65]. SentiStrength is strongly based on LIWC – the work of Pennebaker *et al.*[51] – and therefore also covers the ability to rank the sentiment of grammatically correct text. With reference to the aforementioned criteria:

- Having been designed to rank short informal texts, the SentiStrength system has been shown to consistently outperform a range of competing algorithms[50] in ranking social media vernacular. The SentiStrength package's accuracy is the result of its design criteria of dealing with the colloquial and often grammatically and lexically incorrect nature of the vernacular employed by social media users. It is also capable of assigning sentiment to emoticons[c]; dealing with misspellings; and most importantly dealing with the effects of negation words such as "not" and "never". Since the system builds on the work of Pennebaker *et al.*[51] in the creation of LIWC, SentiStrength is therefore also able to accurately rank formal English text;

- The SentiStrength package is programmatically accessible;

- The SentiStrength package's classification methodology is fully transparent, meaning that the methodology behind determining the sentiment of text can be seen by the user.

Given that SentiStrength was the only classifier which met the aforementioned criteria of: accuracy; speed and convenience of use; and operational transparency, it was chosen as the classification system for this study.

The system has been shown to consistently outperform machine-learning competitors in terms of the accuracy of ranking the sentiment of social media vernacular found on MySpace[d] pages[50]. SentiStrength was tested on a set of 1,041 MySpace comments

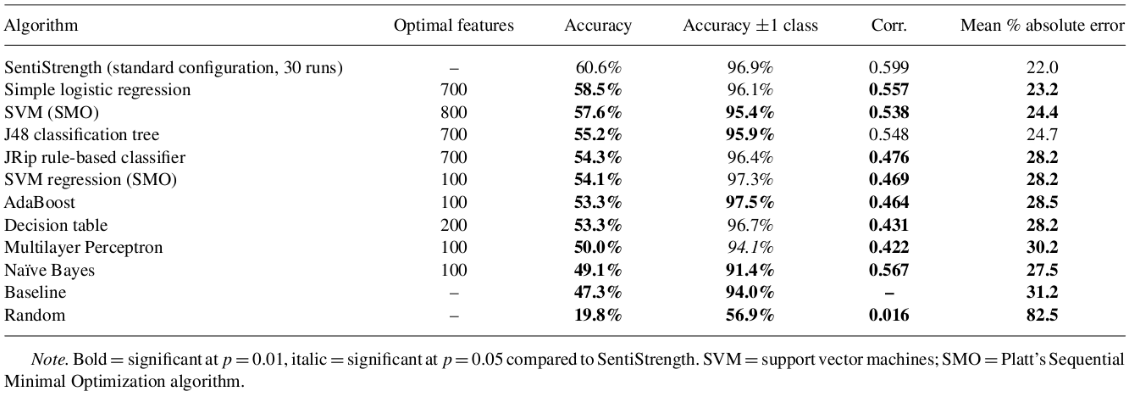
[a] http://sentistrength.wlv.ac.uk/

[b] For example in classifying Tweets relating to the London 2012 Olympics, with the results put up in lights on the EDG Energy London Eye, and classifying the Tweets relating to the 2014 Super Bowl, with the results transformed into a lightshow on the Empire State Building.

[c] A pictorial representation of a human emotion, used in SMS messages as well as in informal internet-based discussions.

[d] MySpace is a music-centric social networking website founded in 2003.

whose sentiments were ranked by three human judges operating independently from a common code book. The system's accuracy was compared to a range of machine-learning classification algorithms used in Weka[a], a popular suite of machine-learning algorithms used for data-mining tasks[66]: Simple logistic regression, Support Vector Machine (Sequential Minimal Optimisation), J48 Classification tree, a JRip rule-based classifier, Support Vector Machine (regression), ADA Boost, Decision table, Multilayer Perceptron, Naïve Bayes as well as random data. When compared to the rankings produced by three human judges, it was shown that SentiStrength's ability to determine the sentiments of MySpace comments was significantly above the best standard machine-learning approaches, as shown below in Table 2 (reproduced from Thelwall *et al.*, 2010), which shows the performance of SentiStrength against the aforementioned machine-learning approaches[66].



Algorithm	Optimal features	Accuracy	Accuracy ±1 class	Corr.	Mean % absolute error
SentiStrength (standard configuration, 30 runs)	–	60.6%	96.9%	0.599	22.0
Simple logistic regression	700	**58.5%**	96.1%	**0.557**	**23.2**
SVM (SMO)	800	**57.6%**	**95.4%**	**0.538**	**24.4**
J48 classification tree	700	**55.2%**	**95.9%**	0.548	24.7
JRip rule-based classifier	700	**54.3%**	96.4%	**0.476**	**28.2**
SVM regression (SMO)	100	**54.1%**	97.3%	**0.469**	**28.2**
AdaBoost	100	**53.3%**	**97.5%**	**0.464**	**28.5**
Decision table	200	**53.3%**	96.7%	**0.431**	**28.2**
Multilayer Perceptron	100	**50.0%**	*94.1%*	**0.422**	**30.2**
Naïve Bayes	100	**49.1%**	**91.4%**	**0.567**	**27.5**
Baseline	–	**47.3%**	**94.0%**	**–**	**31.2**
Random	–	**19.8%**	**56.9%**	**0.016**	**82.5**

Note. Bold = significant at $p = 0.01$, italic = significant at $p = 0.05$ compared to SentiStrength. SVM = support vector machines; SMO = Platt's Sequential Minimal Optimization algorithm.

Table 2: Accuracy of SentiStrength against machine-learning approaches. Reproduced from Thelwall *et al.*, 2010

SentiStrength's accuracy was further demonstrated[67] against the same set of competitive methods and using the same human-based validation procedure, by ranking the sentiments of: YouTube[b] video comments, BBC Forum[c] posts, Digg.com[d] posts, posts on the Runner's World[e] forum, Twitter posts and again MySpace comments. Here, it was determined that SentiStrength exceeds baseline accuracy for negative sentiment strength on all datasets and exceeds baseline accuracy for positive sentiment strength on

[a] http://www.cs.waikato.ac.nz/ml/weka/

[b] YouTube is a video-sharing website founded in 2005. It is the world's third most-visited website.

[c] Discussions of public news as passed through the BBC's online forum, representing serious topics such as national and world news

[d] Discussions of public news as passed through the news and opinion-sharing website http://www.digg.com, representing general news commentary

[e] Runner's World is a global monthly magazine for running enthusiasts founded in 1966.

all datasets except Digg.com and BBC forums, as shown below in Table 3 (reproduced from Thelwall *et al.* 2012), which shows the performance of SentiStrength in ranking texts from the aforementioned internet sources against baseline measures.

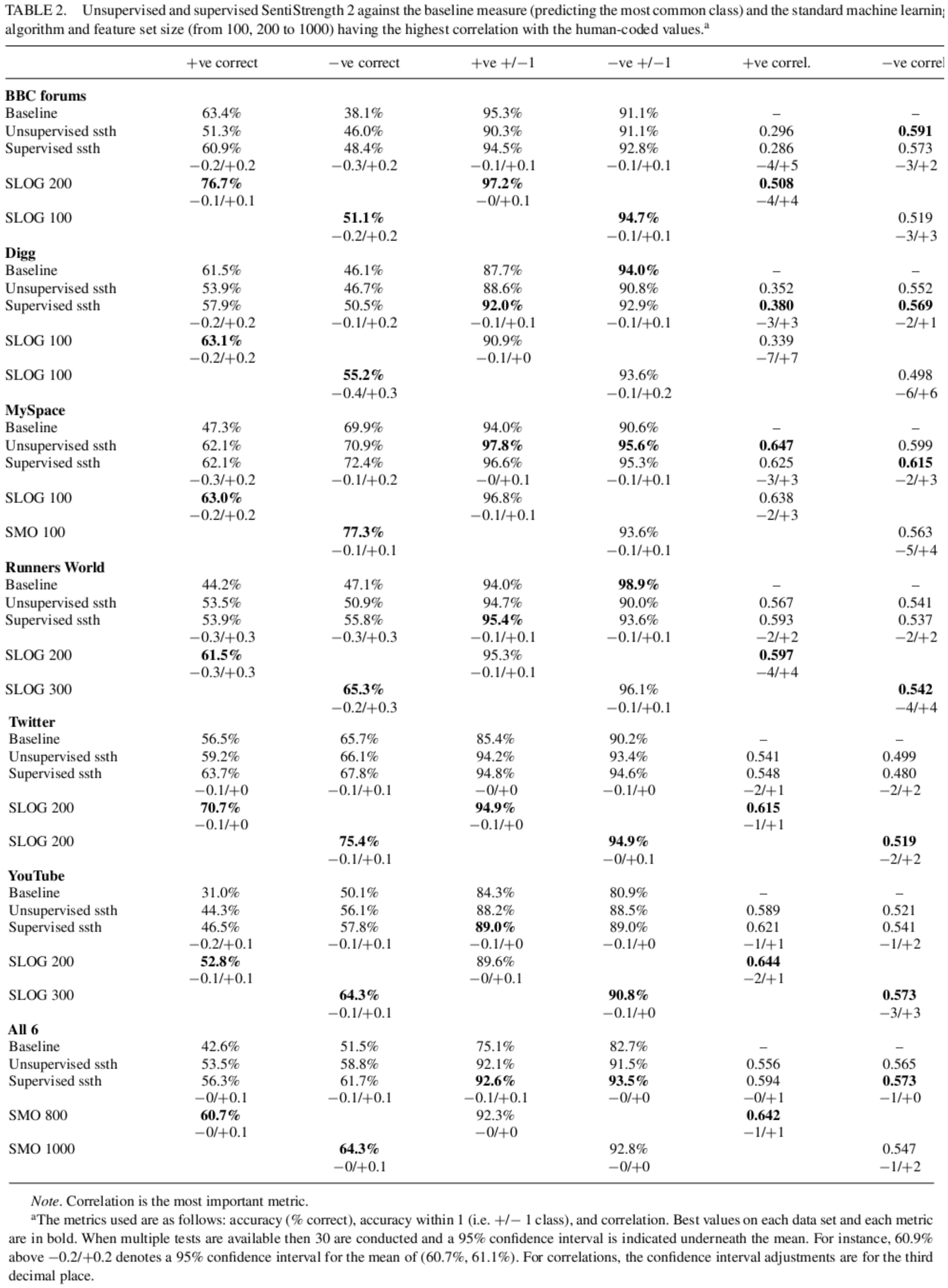


TABLE 2. Unsupervised and supervised SentiStrength 2 against the baseline measure (predicting the most common class) and the standard machine learnin algorithm and feature set size (from 100, 200 to 1000) having the highest correlation with the human-coded values.[a]

	+ve correct	−ve correct	+ve +/−1	−ve +/−1	+ve correl.	−ve correl
BBC forums						
Baseline	63.4%	38.1%	95.3%	91.1%	–	–
Unsupervised ssth	51.3%	46.0%	90.3%	91.1%	0.296	**0.591**
Supervised ssth	60.9%	48.4%	94.5%	92.8%	0.286	0.573
	−0.2/+0.2	−0.3/+0.2	−0.1/+0.1	−0.1/+0.1	−4/+5	−3/+2
SLOG 200	**76.7%**		**97.2%**		**0.508**	
	−0.1/+0.1		−0/+0.1		−4/+4	
SLOG 100		**51.1%**		**94.7%**		0.519
		−0.2/+0.2		−0.1/+0.1		−3/+3
Digg						
Baseline	61.5%	46.1%	87.7%	**94.0%**	–	–
Unsupervised ssth	53.9%	46.7%	88.6%	90.8%	0.352	0.552
Supervised ssth	57.9%	50.5%	**92.0%**	92.9%	**0.380**	**0.569**
	−0.2/+0.2	−0.1/+0.2	−0.1/+0.1	−0.1/+0.1	−3/+3	−2/+1
SLOG 100	**63.1%**		90.9%		0.339	
	−0.2/+0.2		−0.1/+0		−7/+7	
SLOG 100		**55.2%**		93.6%		0.498
		−0.4/+0.3		−0.1/+0.2		−6/+6
MySpace						
Baseline	47.3%	69.9%	94.0%	90.6%	–	–
Unsupervised ssth	62.1%	70.9%	**97.8%**	**95.6%**	**0.647**	0.599
Supervised ssth	62.1%	72.4%	96.6%	95.3%	0.625	**0.615**
	−0.3/+0.2	−0.1/+0.2	−0/+0.1	−0.1/+0.1	−3/+3	−2/+3
SLOG 100	**63.0%**		96.8%		0.638	
	−0.2/+0.2		−0.1/+0.1		−2/+3	
SMO 100		**77.3%**		93.6%		0.563
		−0.1/+0.1		−0.1/+0.1		−5/+4
Runners World						
Baseline	44.2%	47.1%	94.0%	**98.9%**	–	–
Unsupervised ssth	53.5%	50.9%	94.7%	90.0%	0.567	0.541
Supervised ssth	53.9%	55.8%	**95.4%**	93.6%	0.593	0.537
	−0.3/+0.3	−0.3/+0.3	−0.1/+0.1	−0.1/+0.1	−2/+2	−2/+2
SLOG 200	**61.5%**		95.3%		**0.597**	
	−0.3/+0.3		−0.1/+0.1		−4/+4	
SLOG 300		**65.3%**		96.1%		**0.542**
		−0.2/+0.3		−0.1/+0.1		−4/+4
Twitter						
Baseline	56.5%	65.7%	85.4%	90.2%	–	–
Unsupervised ssth	59.2%	66.1%	94.2%	93.4%	0.541	0.499
Supervised ssth	63.7%	67.8%	94.8%	94.6%	0.548	0.480
	−0.1/+0	−0.1/+0.1	−0/+0	−0.1/+0	−2/+1	−2/+2
SLOG 200	**70.7%**		**94.9%**		**0.615**	
	−0.1/+0		−0.1/+0		−1/+1	
SLOG 200		**75.4%**		**94.9%**		**0.519**
		−0.1/+0.1		−0/+0.1		−2/+2
YouTube						
Baseline	31.0%	50.1%	84.3%	80.9%	–	–
Unsupervised ssth	44.3%	56.1%	88.2%	88.5%	0.589	0.521
Supervised ssth	46.5%	57.8%	**89.0%**	89.0%	0.621	0.541
	−0.2/+0.1	−0.1/+0.1	−0.1/+0	−0.1/+0	−1/+1	−1/+2
SLOG 200	**52.8%**		89.6%		**0.644**	
	−0.1/+0.1		−0/+0.1		−2/+1	
SLOG 300		**64.3%**		**90.8%**		**0.573**
		−0.1/+0.1		−0.1/+0		−3/+3
All 6						
Baseline	42.6%	51.5%	75.1%	82.7%	–	–
Unsupervised ssth	53.5%	58.8%	92.1%	91.5%	0.556	0.565
Supervised ssth	56.3%	61.7%	**92.6%**	**93.5%**	0.594	**0.573**
	−0/+0.1	−0.1/+0.1	−0.1/+0.1	−0/+0	−0/+1	−1/+0
SMO 800	**60.7%**		92.3%		**0.642**	
	−0/+0.1		−0/+0		−1/+1	
SMO 1000		**64.3%**		92.8%		0.547
		−0/+0.1		−0/+0		−1/+2

Note. Correlation is the most important metric.

[a] The metrics used are as follows: accuracy (% correct), accuracy within 1 (i.e. +/− 1 class), and correlation. Best values on each data set and each metric are in bold. When multiple tests are available then 30 are conducted and a 95% confidence interval is indicated underneath the mean. For instance, 60.9% above −0.2/+0.2 denotes a 95% confidence interval for the mean of (60.7%, 61.1%). For correlations, the confidence interval adjustments are for the third decimal place.

TABLE 3: ACCURACY OF SENTISTRENGTH AGAINST BASELINE COMPETITORS. REPRODUCED FROM THELWALL *ET AL.*, 2012

Thus, SentiStrength performed significantly above[67] its competitors for ranking of sentiment across six social web datasets which differ substantially in origin and content.

By considering the demonstrated accuracy of the SentiStrength package in ranking the sentiments of short informal texts found on the social web, and the unsuitability of the leading research-oriented and commercial-oriented sentiment classifiers to ranking such language, it was possible to identify SentiStrength as the tool best-suited to the study. Thus, as discussed earlier in this Thesis, the final version of the TCF framework was developed in such a manner that incoming Tweets were parsed by SentiStrength at the point of collection.

In the study's implementation of SentiStrength, for each incoming Tweet, the TCF stored the date/time-stamp of creation, and SentiStrength's sentiment outputs which consisted of a positive sentiment score (i.e., how positive a string of text is) and a negative sentiment score[a] (i.e., how negative a string of text is). Positive sentiments are ranked on a scale of +1 (least positive) to +5 (most positive); and negative sentiments are ranked on a scale of -1 (least negative) to -5 (most negative). With the addition of the positive and negative sentiments for a given string, it is also possible to determine the text's overall net sentiment. This is calculated by subtracting the negative sentiment from the positive sentiment for each message. The resultant net score is ranked on a scale of -4 (most negative) through 0 (average) to +4 (most positive).

The following is an example of a string of text containing both a positive and a negative sentiment component, as ranked by SentiStrength:

"I love David Cameron, but hate the current political climate"

SentiStrength ranks this string as having a positive sentiment score of +3 (on a scale of +1 to +5), and a negative sentiment score of -4 (on a scale of -1 to -5). This gives a net sentiment score of -1 on a scale of -4 (most negative) to +4 (most positive).

[a] SentiStrength can be configured to provide a combined net sentiment for a given input string. However, the segregation of a string's sentiment scores into separate 'positive' and 'negative' components provides an additional dimension to the dataset. In this study, all three sentiment dimensions provided by SentiStrength were investigated.

5.2 Data sample

It is important to consider the chronological frame of reference in order to obtain a representative and viable data sample for the study. The following criteria must be considered in determining the sample.

First, literature indicates the intuitive fact that annual stock-market volatilities are influenced by long-term global macroeconomic trends[68,69]. Therefore, the chronological period considered must be selected to minimise the effects of routine quarterly updates[a] of ever-changing macroeconomic trends, whilst still offering a range of intra-day market volatilities. Furthermore, the data-set must be sufficiently small to minimise the effects of seasonality[b] (as discussed in depth in Chapter 5.6.2.1). The effect of seasonality in social media data cannot be mapped due to the inherent unavailability of historic datasets, and therefore cannot be removed accurately based on the analysis of such historic data.

As well as selecting a time-period that minimises variability of the economic data, it is important to be aware of changes to the product supplying the raw data. Twitter is a commercial entity that routinely updates its products to maintain competitiveness against its rivals such as Facebook. These changes can dramatically alter the core Twitter product, and therefore either alienate its customer base, or attract new users – resulting in changes to the nature of the content of the data that the network transmits. For example, in April 2014, Twitter announced a substantial redesign to its graphical user interface[70], resulting in comments that the network was mimicking Facebook's design[71]. This was interpreted by users as a phasing out of the popular feature known hashtags[c]. Considering that Facebook is known to experience dramatic changes to the demographical makeup of its users on an annual basis[46], it is reasonable to state that dramatic alterations to Twitter's core product (such as a major redesign of the graphical user interface, or the intended removal of hashtags) will influence the consistency of the Tweet data used by the study, driven the resultant changing demographics of Twitter's users.

[a] Macroeconomic data is typically reported on a quarter-by-quarter basis. With reference to this study, the United States Department of Commerce Bureau of Economic analysis, and the UK's Bank of England report macroeconomic data on a quarterly basis.

[b] Seasonality is the effect in time-series data that is driven by economic cycles influenced by the time of year.

[c] A hashtag is an unspaced string prefixed with the "#"sign, which is used on Twitter to tag Tweets according to a particular topic. It is a popular method for Tweet promotion on the network. For example, a Tweet tagged with the "#OccupyWallStreet" hasthag would denote that message's affiliation with the anti-consumerist protest movement of late 2011.

For these reasons, only a 3-month data-sample is considered in this study. The collection period lasted three months, from 11th December 2012 to 12th March 2013, and resulted in 4.71GB of raw time-stamped sentiments, equating to 112,628,180 rows of data. The collection methodology is further discussed in Chapter 4. As discussed in Chapter 4.3, the financial data consisted of Futures[a] price data obtained from Fulcrum Asset Management[b], and CFD[c] price data obtained from the Swiss foreign-exchange bank and marketplace Dukascopy[d]. Including this financial data, this 3-month dataset consisted of 451,653,196 raw data-points. While the collection time is relatively small, the quantity of sentiment and related financial data is large. This provides an indication of the data processing requirements for diminishing asset price uncertainty with social media data, and the drive behind the methodological decisions made herein. Furthermore, past studies in this space do not stipulate a minimum chronological data-size as it is specific to each study – indeed one past work on the analysis of Tweet message sentiments and volumes considered just a 32-day dataset[49].

The 3-month chronological period encompassed a range of holiday periods and normal-activity periods for the UK and the US financial markets (as shown in Table 4), resulting in a spectrum of market conditions whilst only encompassing a single macroeconomic data update: the Q4-2012 to the Q1-2013 transition period[72, 73], meeting the first criteria of minimised macroeconomic and seasonality trends.

Period	US Financial Market Holiday	UK Financial Market Holiday
25 December 2013	Christmas	Christmas
26 December 2013	-	Boxing Day
01 January 2013	New Year's Day	New Year's Day
21 January 2013	Martin Luther King Jr. Day	-
18 February 2013	President's Day	-

Table 4: Financial market holiday periods considered by the study

Table 4 shows that the 3-month chronological data-sample considered by the study encompassed four days of US market-closure, and three days of UK market-closure, and

[a] In finance, a Futures Contract is a standardised financial derivatives contract describing the intended purchase or sale of a financial instrument, at a pre-determined future date and price. It is used as a method for hedging a financial investment position, or speculating on the price movement of an underlying asset.

[b] http://www.fulcrumasset.com/

[c] Contracts for Differences (CFDs) are arrangements in a Futures contract which describe the delivery of cash payments between a buyer and seller equating to the difference between the current value of an asset and its value at a contract time. They are financial derivatives which allow market traders to benefit from changes in prices of the underlying asset, without owning the asset itself.

[d] http://www.dukascopy.com/

therefore providing a range of market conditions for the study. This period also met the second criteria, having the effect of avoiding encapsulating any major redesigns or alterations to Twitter's product. Indeed, the period from 11th December 2012 to 12th March 2013 did not include any major Twitter product releases or alterations[74-77] other than minor improvements to the visibility of rich media on the network.

5.3 Time-series dependency measures

The social media and financial data must be arranged in a manner suitable for assessing their interdependencies. At the point of output from the TCF, the social media data are continuous, meaning that the data have a particular value for only an infinitesimally short amount of time and that any number of data-points can exist in a time-period[78]. The TCF can therefore produce any number of Tweet sentiments for any given time-frame. In contrast, the financial data used by this study are discrete[a], meaning that the data values occur at separate and distinct points in time as a result of sampling into time-windows of a desired size[78]. The datasets are therefore not yet arranged in a manner suitable for assessing their dependencies.

To allow for the comparison of the Twitter data to the financial data, what is needed is a method for standardisation of the two datasets to identical discretisation levels. In the financial services industry, the choice of discretisation frequency is often ad-hoc, typically dictated by the observation intervals of the available data[79]. As discussed in Chapter 4.1, the development of SocialSTORM[57] provided preliminary access to Twitter data for initial exploration of the relationships between social media data and financial data. Whilst the Twitter data provided by SocialSTORM was continuous, as is the case with the TCF, the financial data used during this preliminary investigation was not available to resolutions smaller than hourly[b80]. These preliminary investigations were therefore performed on social media and financial datasets discretised to hourly windows, showing support for the existence of dependencies between the two[80-82].

Due to this past data limitation and that in the financial services industry the choice of discretisation frequency dictated by the observation intervals of the available data, this

[a] As provided by the data providers: Fulcrum Asset Management and Dukascopy

[b] Financial data used for the preliminary investigation was sourced from Thomson Reuters and from Fulcrum Asset Management, and was discretised to hourly windows due to the unavailability of higher-resolution data at the time.

study's primary level of discretisation is hourly. However, the robustness of the relationships at different discretisation levels is tested, as discussed in Chapter 7.1.

Next, it should be noted that message sentiments, their volumes, and asset prices are not in a static steady-state, and are instead time-dependent (i.e., dynamic). What is therefore needed is a conversion of these dynamic time-series to time-independent changes in their states. To achieve this, part of the study's data analytics process involved the calculation of changes in the social media and financial time-series between adjacent data-discretisation windows.

To satisfy the aforementioned criteria of converting the study's datasets into discrete static variables, the social media data and the corresponding financial data for each Twitter filter were first discretised by way of arithmetic mean averaging into discretised non-overlapping consecutive windows. As discussed above, these windows were of 1-hour in size, on the hour – i.e., the discretised adjacent windows are placed on the hour. The robustness of the results is tested against variation in window size is detailed in Chapter 7.1. The robustness of the results is also tested in Chapter 7.2 against variation in the offset of the adjacent discretisation windows – i.e., the discretised windows adjacent are not placed on the hour. The discretisation procedure was performed as follows:

1. A discretised time-series T of time-stamps with elements $\mathrm{T_i}$ is created, where $\mathrm{T_1}$ = 00:00:00 on 11th December 2012 and concluding at 23:59:59 on 11th March 2013 (bringing the data-capture period up to 12th March 2013, giving a total of 90 days).

2. The number of periods per 24-hours is determined as a function of the desired window size, W when expressed in hours:

$$\mathrm{N_{periods}} = \frac{24}{\mathrm{W}}$$

3. The number of elements in the discretised time-series T is therefore:

$$\mathrm{T_n} = \mathrm{N_{periods}} \times 90$$

4. It is then identified whether the input data time-series of price, sentiment and message volume, $\mathrm{I_{price}}, \mathrm{I_{sentiment}}, \mathrm{I_{message\ volume}}$ belong to each location in the discretised time-series T. An input data-point I is deemed to belong to a location

in the discretised time-series T if its time-stamp is between up to and including the time-stamp for the current location in the discretised time-series, $\mathrm{T_i}$, and above but not including the time-stamp for the chronologically previous location in the discretised time-series, i.e., $\mathrm{T_{i-1}}$.

5. For each location in the discretised time-series T, the discretised means of the values for each of the corresponding input data series of price, sentiment and message volume, $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$ are determined. Denoted $\overline{\mathrm{D_{price_{T_n}}}}$, $\overline{\mathrm{D_{sentiment_{T_n}}}}$ and $\overline{\mathrm{D_{message\ volume_{T_n}}}}$ respectively, these are calculated as:

$$\overline{\mathrm{D_{price_{T_1}}}} = \frac{\mathrm{I_{price_1}} + \mathrm{I_{price_2}} + \cdots \mathrm{I_{price_n}}}{\mathrm{n}}$$

$$\overline{\mathrm{D_{sentiment_{T_1}}}} = \frac{\mathrm{I_{sentiment_1}} + \mathrm{I_{sentiment_2}} + \cdots \mathrm{I_{sentiment_n}}}{\mathrm{n}}$$

$$\overline{\mathrm{D_{message\ volume_{T_1}}}} = \frac{\mathrm{I_{message\ volume_1}} + \mathrm{I_{message\ volume_2}} + \cdots \mathrm{I_{message\ volume_n}}}{\mathrm{n}}$$

6. Finally, the changes in these discretised mean values of $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$ are then calculated to produce static variables. Denoted $\Delta\overline{\mathrm{D_{price_{T_n}}}}$, $\Delta\overline{\mathrm{D_{sentiment_{T_n}}}}$ and $\Delta\overline{\mathrm{D_{message\ volume_{T_n}}}}$ respectively, these are calculated as:

$$\Delta\overline{\mathrm{D_{price_{T_1}}}} = \overline{\mathrm{D_{price_{T_1}}}} - \overline{\mathrm{D_{price_{T_{1-1}}}}}$$

$$\Delta\overline{\mathrm{D_{sentiment_{T_1}}}} = \overline{\mathrm{D_{sentiment_{T_1}}}} - \overline{\mathrm{D_{sentiment_{T_{1-1}}}}}$$

$$\Delta\overline{\mathrm{D_{message\ volume_{T_1}}}} = \overline{\mathrm{D_{message\ volume_{T_1}}}} - \overline{\mathrm{D_{message\ volume_{T_{1-1}}}}}$$

In this manner, this methodology also normalises the data by the volume of data-points for each element in the time-series T.

7. Note, the values of $\Delta\overline{\mathrm{D_{price_{T_1}}}}$, $\Delta\overline{\mathrm{D_{sentiment_{T_1}}}}$ and $\Delta\overline{\mathrm{D_{message\ volume_{T_1}}}}$ (i.e., for element $\mathrm{T_1}$) are empty as these are the first entries in the discretised time-series T and therefore there are no prior elements from which to calculate the changes in these discretised mean values of $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$.

This study therefore measures the dependency between the discretised values of $\Delta_{\text{sentiment}}$ vs. the Δ_{price}, and $\Delta_{\text{message volume}}$ vs. the Δ_{price} for each financial-instrument/Twitter-Filter combination.

What is therefore needed is a measure of dependency which allows for the assessment of the extent to which this social media data leads the financial data. By mirroring past works in this space[7, 11, 12, 37, 49], the present study first considered *linear regression analysis*, identifying its limitations in suitability to the assessment of the study's datasets. The limitations of using linear regression analysis for the study's dataset were mitigated by next using *information theory* as the measure of dependency. This is discussed in the next two chapters.

5.3.1 Linear regression analysis

The statistical relationship between any two random variables or sets of data can be evaluated using Correlation analysis – a broad class of statistical methods for observing the inter-dependence of variables, a form of which is Pearson's r. Developed from Francis Galton's late 19th Century work on correlation[83] by Karl Pearson in the early 20th Century, it is suitable for determining the extent to which a relationship between two variables can be approximated by a linear relationship. It is a measure of linear dependence between two variables[84] and is employed in various realms from finance to engineering.

Pearson's r is a measure of the covariance[a] between two variables divided by the product of their standard deviations. The resultant value ranges from -1 (for a strong negative correlation) to +1 (for a strong positive correlation).

For a sample, Pearson's r is given as:

$$\mathrm{r} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i} - \overline{\mathrm{X}})(\mathrm{Y_i} - \overline{\mathrm{Y}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i} - \overline{\mathrm{X}})^2}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \overline{\mathrm{Y}})^2}}$$

Where:

- X refers to Variable 1. $\mathrm{X_i}$ refers to the ith value within the series Variable 1;
- Y refers to Variable 2. $\mathrm{Y_i}$ refers to the ith value within the series Variable 2;

[a] Covariance is a measure of how much two random variables change together.

- $\overline{X}$ refers to the arithmetic mean of the values in series Variable 1;
- $\overline{Y}$ refers to the arithmetic mean of the values in series Variable 2;
- The numerator refers to the covariance between variables X and Y;
- The denominator refers to the product of the standard deviations for a sample of variables X and Y.

This study investigates linear regression analysis as a measure of dependency to mirror past works[7,11,12,37,49] which have also used this measure.

The implementation of linear regression analysis is discussed in Chapter 5.4.

5.3.2 Information theory

Linear regression analysis is limited by the assumption that the nature of the relationship being investigated is linear[85-87]. Considering the common recognition of the non-linearity of financial time-series[88,89], an alternative measure of dependency is needed in order to not adhere to the assumption of linearity within the study's dataset.

Multi-order or non-linear analyses[90] can be used if the data being investigated cannot be approximated by a linear model, or if the assumption that the relationships are linear cannot be made. These alternative methodologies typically require approximation of the datasets using some form of model – a process which replaces the underlying raw data itself with an approximation of the raw data. The approximation of any data removes potentially-valuable detail from the data. Therefore, what is needed is a measure of dependency which does not require the assumption that the underlying relationships are linear, or the approximation of the raw data with a descriptive model.

Statistical-analysis constructs exist which allow for the relationships between time-series to be established without needing to know the data's mean-variance or probability distribution characteristics[85], or without approximation of the raw data with a descriptive model. One such construct – information theory – refers to a branch of applied mathematics centred on the quantification of information. Based on probability theory, the construct has found use in applications requiring signal processing and statistical inference in areas such finance and engineering[91]. It is employed in the present study as a measure of dependency between the social media data and the financial data, thus overcoming the limitations of linear regression analysis without requiring the approximation of the raw data with a descriptive model.

A key measure used in information theory is entropy[a], which quantifies the uncertainty involved in predicting the value of a random variable, and has been strongly defended as having a relationship with predictability and dependence[92]. The entropy H of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X.

If $\mathbb{X}$ is the set of all messages $\{x_1, \dots, x_n\}$ that X could be, and $p(x)$ is the probability of some x ∈ $\mathbb{X}$, then the entropy H of X is defined as:

$$H(X) = \mathbb{E}_X[I(x)] = -\sum_{x \in \mathbb{X}} p(x) \log p(x)$$

Where:

- $\mathbb{X}$ is the set of all messages $\{x_i, \dots, x_n\}$ that X could be;
- $p(x)$ is the probability of some x ∈ $\mathbb{X}$;
- $I(x)$, the self-information, is the entropy contribution of an individual message. It is a measure of the information content associated with the outcome of a random variable, and is dependent on the probability of that event. The smaller its probability, the larger the self-information associated with receiving the information that event occurred. For a probabilistic event, the self-information $I(\omega_n)$ associated with outcome ω_n with probability $P(\omega_n)$ is defined as:

$$I(\omega_n) = \log\left(\frac{1}{P(\omega_n)}\right) = -\log\big(P(\omega_n)\big)$$

- $\mathbb{E}_X$ is the expected value.

This quantification of information is applied to the measure of variables' mutual dependence, known as *mutual information*[b]. This is a measure of the amount of

[a] Referring to the Shannon entropy, measured in bits, which quantifies the expected value of the information contained in a message

[b] Measured in bits, the measure of Mutual Information quantifies the information that two random variables share

information which can be obtained about one random variable by observing another[93], measuring how much knowing one variable reduces the uncertainty about the other. If two random variables, X and Y are independent, then observing X reveals no information about Y, and the mutual information is zero. Conversely, if X and Y are fully deterministic about one another (a special case where two random variables are identical), then the mutual information is the same as the uncertainty (i.e., the entropy, defined above) contained in Y or X.

Mutual information is used as a measure of the dependency between the discretised changes in the Twitter data: $\Delta_{\mathrm{sentiment}}$ or $\Delta_{\mathrm{message\ volume}}$, and changes in the financial data: Δ_{price}. The greater the mutual information between the changes in the Twitter data and the changes in the financial data, the more we can establish about the nature of the financial data by observing the Twitter data.

The mutual information of a discrete random variable X based on the observation of a discrete random variable Y is given by:

$$\mathrm{I(X;Y)} = \sum_{\mathrm{y \in Y}} \sum_{\mathrm{x \in X}} \mathrm{p(x,y)log}\left(\frac{\mathrm{p(x,y)}}{\mathrm{p(x)p(y)}}\right)$$

Where:

- $\mathrm{p(x,y)}$ is the joint probability distribution function of X and Y, which describes the probability that each of $\mathrm{X,Y}$ falls in a range of values specified for that variable. In the present study, the joint probability distribution $\mathrm{p(x,y)}$ is identified using a bivariate three-dimensional histogram[94], calculated using MATLAB's `hist3`[a] bivariate histogram function. Therefore, whilst the ranges of the two distributions are not identical, a common number of bins is used for both. This therefore results in non-identical bin widths determined separately for each distribution;

- $\mathrm{p(x)}$ is the marginal probability distribution function of X, which is the probability distribution of the values contained within the subset of X without reference to the values of other variables;

[a] http://www.mathworks.co.uk/help/stats/hist3.html

- $p(y)$ is the marginal probability distribution function of Y, which is the probability distribution of the values contained within the subset of Y without reference to the values of other variables.

Note that the measure of mutual information is symmetric, i.e., $I(X; Y) = I(Y; X)$. Therefore, an additional test is needed to ascertain whether the social media data is more proactive than reactive, relative to the financial data. This test is discussed in Chapter 5.5.2.

5.3.2.1 Estimating probability distributions with binning

Since the computation of entropy, which is necessary as part of the process for calculating mutual information, is based on the probability of the values within the dataset being investigated, it is necessary to estimate their probability distributions. In this study, such probability distributions are estimated using a histogram. The selection of histogram bin sizes is performed using the Sturges' histogram rule[95], a well-documented and often-used method, frequently found as the default tool for histogram binning in statistical packages[96]. In addition, as detailed in Chapter 7.3, the results of the study are tested against another well-known histogram-estimation method to demonstrate the suitability of histogram selection using this rule. It is defined as:

$$\omega = \frac{r}{1 + \log_2(n)}$$

Where:

- r is the range of values within the dataset;
- n is the number of elements in the dataset;
- ω is the ideal bin width to be used for the histogram;
- Calculating r/ω gives the number of bins for the dataset.

5.3.2.2 Time-shifted mutual information comparisons

By quantifying the mutual information between the social media data and the financial data at different time-shifts[a], it is possible to evaluate how much information Twitter data contains about the returns of financial markets ahead of time. For social media data to lead the financial markets ahead of time, the quantity of mutual information between the Twitter data and the financial data must be greater at a chronologically leading time-shift between the two datasets than at no time-shift. The scope of this study is the exploration of social media's ability to lead financial data up to 24-hours ahead of time.

The use of information theory in this study is not specifically designed to extract or maximise positive relationships from the study's dataset. Rather, this measure of dependency, which is not specifically designed for social media data, is implemented to conservatively identify when and if social media can lead financial data. This is to ascertaining if a conservative measure of dependency that is not specifically tailored to the underlying dataset can identify instances of social media leading financial data. If so, these findings would give further support and credence to recent academic exercises[13-15] in portfolio structuring based on the analysis of social media data using retroactively-calculated profit maximisation as the success criterion.

The implementation of information theory is discussed in Chapter 5.5.

[a] The 'time-shift' is an artificial chronological off-set between the two time-series such that one leads the other.

5.4 Implementation of linear regression analysis

Since linear regression analysis calculations are straightforward, as described in Chapter 5.3.1), this measure of dependency was implemented in three separate experiments to determine its efficacy in a range of configurations. These experiments used three different binning[a]97 methods:

1. The study's dataset was evaluated as one time-series using no binning;
2. The study's dataset was subdivided into bins to segregate the social media data and financial data based on a method which identifies instances of non-zero financial trading volume;
3. The study's dataset was subdivided into bins to segregate the social media and financial data based on a method which identifies instances of non-zero financial returns. This results in data bins which only encapsulate instances where the financial securities produced a non-zero return.

For each of the grouping methods described above, the simple conditions of: no time-shift and 24-hour backward looking SMA smoothing and hourly discretisation of the data were used to assess if there are correlations between the hourly changes in the social media sentiment and the securities' hourly returns at no chronological lag. With regards to the three binning methods:

- Experiment 1: In the case of the first grouping method, the social media data were regressed against the financial data. The results of this experiment are detailed in Chapter 6.2.1.
- Experiment 2: In the case of the second grouping method, the social media data were regressed against the financial data for each bin of non-zero trading volumes. The results of this experiment are detailed in Chapter 6.2.2.
- Experiment 3: In the case of the third grouping method, the social media data were regressed against the financial data for each bin of non-zero returns. The results of this experiment are detailed in Chapter 6.2.3.

[a] Binning is a data-splitting technique. In this application, it is used for data reduction. The study's dataset is reduced in size according to a set of criteria, producing sub-datasets which match required conditions.

5.5 Implementation of information theory

5.5.1 The information surplus evaluation metric

A mutual information-based evaluation metric was developed for this study which allows for the calculation of the extent to which the changes in the sentiments of social media messages or the changes in their volumes contain statistically-significant lead-time information about financial market returns. Specifically, the changes in the sentiments and message volumes of Tweets from the USA and the UK filtered using forty-four specifically-tailored Twitter filters (as listed in Table 1) were evaluated against the returns of twenty-eight financial instruments collected over the 3-month period from 11th December 2012 to 12th March 2013[a]. Combinations of Twitter filters with their corresponding financial data are referred to as 'financial-instrument/Twitter-Filter combinations'.

The mutual information[85] between the two time-series at different time-shifts was considered. A time-shift is an artificially-instated chronological offset between the two time-series. Since mutual information shows the amount of uncertainty in a time-series which can be removed by observing another time-series, it is possible to quantify the extent to which changes in Twitter sentiment or message volumes can remove the uncertainty about the future returns of financial assets by instituting a range of time-shifts between the social media and financial returns time-series.

Therefore, for each financial-instrument/Twitter-Filter combination, the mutual information between changes in the social media data and changes in the corresponding financial data at no time-shift (when social media data and financial data are chronologically superimposed) is first determined. Based on the justifications given in Chapter 5.2, the study considers hourly changes in the dataset – however the robustness of the results are checked against the changes in different discretisation window sizes in Chapter 7.1.

Next, a leading time-shift is instituted between the two time-series, such that the social media data precedes the financial data, and determine the amount of mutual information now available compared to the condition where the time-shift between the two time-series was zero.

[a] See Chapter 5.2 for explanations for the selection of this time range.

Suppose that the amount of mutual information μ between the social media data and financial data at a time-shift of zero hours $\mathrm{L} = 0$ is equal to x:

$$\mu_{\mathrm{L}=0} = \mathrm{x}$$

Now, suppose that the amount of mutual information μ between the social media data and financial data at a leading time-shift of $\mathrm{L} > 0$ is equal to y:

$$\mu_{\mathrm{L}>0} = \mathrm{y}$$

The percentage increase in mutual information between the two aforementioned conditions, $\mu_{\%\mathrm{inc}}$, from $\mu_{\mathrm{L}=0} = \mathrm{x}$ to $\mu_{\mathrm{L}>0} = \mathrm{y}$, is referred to as the *information surplus*. If the information surplus is positive, i.e., $\mu_{\%\mathrm{inc}} > 0$, then the social media data contains more mutual information about financial data at a leading time-shift of $\mathrm{L} > 0$ than at no time-shift, $\mathrm{L} = 0$. In such a scenario, the social media data contains lead-time information about financial data as it removes more uncertainty, ahead of time, about the financial data time-series than at no leading time-shift. Conversely, if the information surplus is negative, i.e., $\mu_{\%\mathrm{inc}} < 0$, then the social media data contains less mutual information about financial data at a leading time-shift of $\mathrm{L} > 0$ than at no time-shift, $\mathrm{L} = 0$. In such a scenario, the social media data does not contain lead-time information about financial data as it removes less uncertainty, ahead of time, about the financial data time-series than at no leading time-shift.

The social media data are offset ahead of the financial data from 0-hours to 24-hours in 1-hour increments[a]. The aforementioned mutual information calculations are then performed on the social media data (for hourly changes in all three sentiment types: positive; negative; and net, and for hourly changes in the message volumes) and financial data from all forty-four Twitter filters considered in this study. In this manner the information surplus is determined for each financial-instrument/Twitter-Filter combination. This allows for the identification of the leading time-shift(s), if any, at which the social media data leads the financial data. Finally, the sentiment type (positive; negative; or net) which results in the maximum information surplus for each financial-instrument/Twitter-Filter combination is identified, and whether the sentiment data outperforms message volumes in leading the securities' returns.

[a] The robustness of the results based on this offset window size is explored in Chapter 7, in which different offset window sizes are considered.

5.5.2 Does social media data lead or trail financial data?

The study's aim is to determine for which assets do the hourly changes in social media data lead securities' hourly returns in a statistically-significant manner. Therefore, firstly it must be ascertained that the information surplus methodology is able to identify financial instruments for which the social media data carries more information about the financial data before price changes rather than after price changes. In such a manner, the notion that social media data contains leading information about financial data rather than merely reacting to it can be supported. To do this, for each time-shift offset of 1-hour to 24-hours such that the social media data leads the financial data, the mutual information between the two time-series of 90 days of data[a] is calculated, thus identifying the 'per-time-shift leading mutual information' for each financial-instrument/Twitter-Filter combination. Then the 'mean trailing mutual information' is determined: the mean mutual information between the social media data and the financial data for each financial-instrument/Twitter-Filter combination when offsetting the two time-series so that the social media data trails (rather than leads) the financial data. An example of this is reported in Figure 8. In such a manner it is possible identify instances when for a given leading time-shift between social media data and financial data, the social media data is more leading than trailing. For a given leading time-shift, the study only permits those financial-instrument/Twitter-Filter combinations for which the per-time-shift leading mutual information exceeds the mean trailing mutual information – thus identifying that the social media leads rather than trails the financial data.

[a] See Chapter 5.2 for an explanation of the length of the time-series

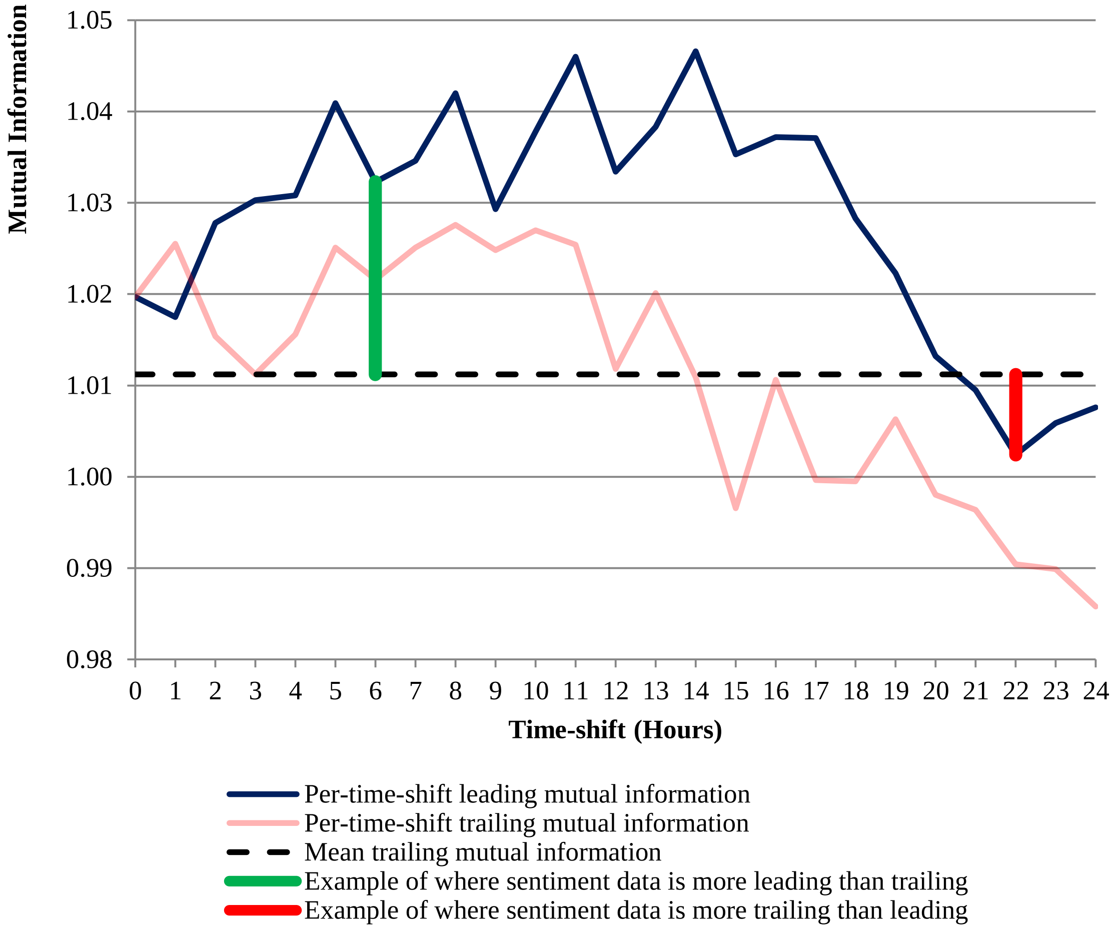


FIGURE 8: EXAMPLE SHOWING IF HOURLY CHANGES IN TWEET SENTIMENTS ARE MORE LEADING THAN TRAILING RELATIVE TO A SECURITY'S HOURLY RETURNS

By way of example, Figure 8 demonstrates the mutual information between hourly changes in sentiment data for the Twitter Filter: "$GOOG" AND/OR "Google" compared with the hourly returns of Google, Inc. CFDs. This example only considers the changes in the negative sentiments as calculated by SentiStrength. Here, the data are presented for time-shifts between 0 and 24-hours both in a leading configuration (such that changes in the sentiment data lead the returns) and in a trailing configuration (such that the returns lead the changes in the sentiment data). The study only permits those time-shifts for which the per-time-shift leading mutual information exceeds the mean trailing mutual information, as indicated by the vertical green bar, and reject those time-shifts for which per-time-shift leading mutual information is less than the mean trailing mutual information, as indicated by the vertical red bar. This process identifies the time-shifts for which the social media leads rather than trails the financial data.

Next, the information surplus is calculated for each such leading time-shift relative to no time-shift, and thus the study only permits those time-shifts which show a positive information surplus, as illustrated by way of example in Figure 9.

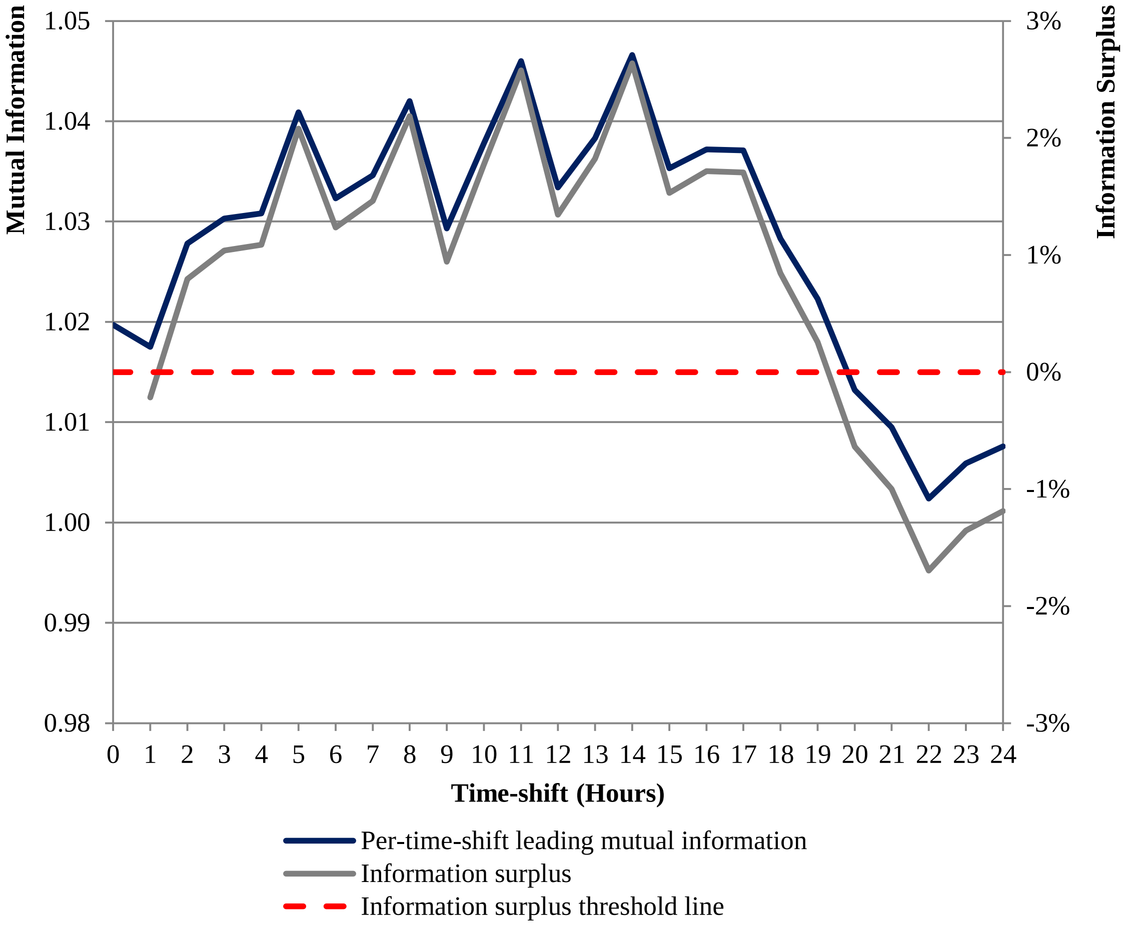


FIGURE 9: EXAMPLE SHOWING IF HOURLY CHANGES IN TWEET SENTIMENTS CAN LEAD A SECURITY'S HOURLY RETURNS

By way of example, Figure 9 demonstrates the information surplus between hourly changes in the sentiment data for the Twitter Filter: “$GOOG” AND/OR “Google” and the hourly returns of Google, Inc. CFDs. As in Figure 8, this example only considers the negative sentiments generated by SentiStrength for this financial-instrument/Twitter-Filter combination. The ‘Information surplus threshold line’ is included only for visual clarity as it visually identifies the percentage increase level of 0% in the information surplus for time-shifts > 0 hours, relative to the information surplus for a time-shift of zero hours. This line is of importance: for the changes in the social media data to be considered leading, they must demonstrate positive information surplus values at time-shifts where the hourly changes in the social media time-series are offset such that they lead the financial data. As in this example, the study only permits those leading time-shifts for which the information surplus curve is above this information surplus threshold line of 0%.

To summarise, this filtering mechanism identifies instances when changes in social media data carry more information about a security's hourly returns ahead of time than at zero leading time-shift to show which time-shifts, if any, result in the social media data preceding the financial data in a manner such that it is more leading than trailing. A negative information surplus would imply that sentiment data carries less information about financial data than at no time-shift between the social media and financial data time-series.

5.5.3 Testing for statistical significance

The final task is to determine the statistical-significance of instances where the social media data are shown to be more leading than trailing for a given time-shift. To achieve this, the hourly changes in social media data (for message volumes this is: $\Delta_{\text{message volume}}$, and for each sentiment type this is: $\Delta_{\text{sentiment}}$) are randomly permutated 10,000 times with respect to the financial data: Δ_{price}. This allows for the calculation of the randomised mutual information at each permutation for a given financial-instrument/Twitter-Filter combination for each leading time-shift from 0 hours to 24-hours. This therefore allows for the calculation of the frequency at which the observed mutual information between the social media data and the financial data exceeds the randomised mutual information over the 10,000 random permutations. The observed mutual information for each sentiment type (positive, negative or net) is evaluated against the randomised mutual information for each sentiment type independently to avoid a multiple-hypothesis testing configuration. The study therefore admits those leading time-shifts for which the observed mutual information between the social media data and the financial data is greater than the randomised mutual information with a statistically-significant confidence interval of 99%. Note, in order to echo recent studies which evaluate Google Search Trends[13-15] and Yahoo! search engine data[16] message volumes against financial market performance, the tests for statistical-significance are also repeated to evaluate the extent to which hourly changes in Tweet message volumes $(\Delta_{\text{message volume}})$ lead absolute hourly changes in securities' prices $(|\Delta_{\text{price}}|)$.

In summary, by satisfying the aforementioned caveats the study tests whether changes in social media sentiments and/or message volumes lead securities' returns; whether changes in social media sentiments and/or message volumes are more leading than trailing when evaluated against hourly financial returns at different time-shifts; and then the resultant relationships are tested for statistical-significance. Consequently, the study identifies statistically-significant leading time-shifts for which hourly changes in the sentiments and/or message volumes lead the securities' hourly returns, an example of which is shown in Figure 10.

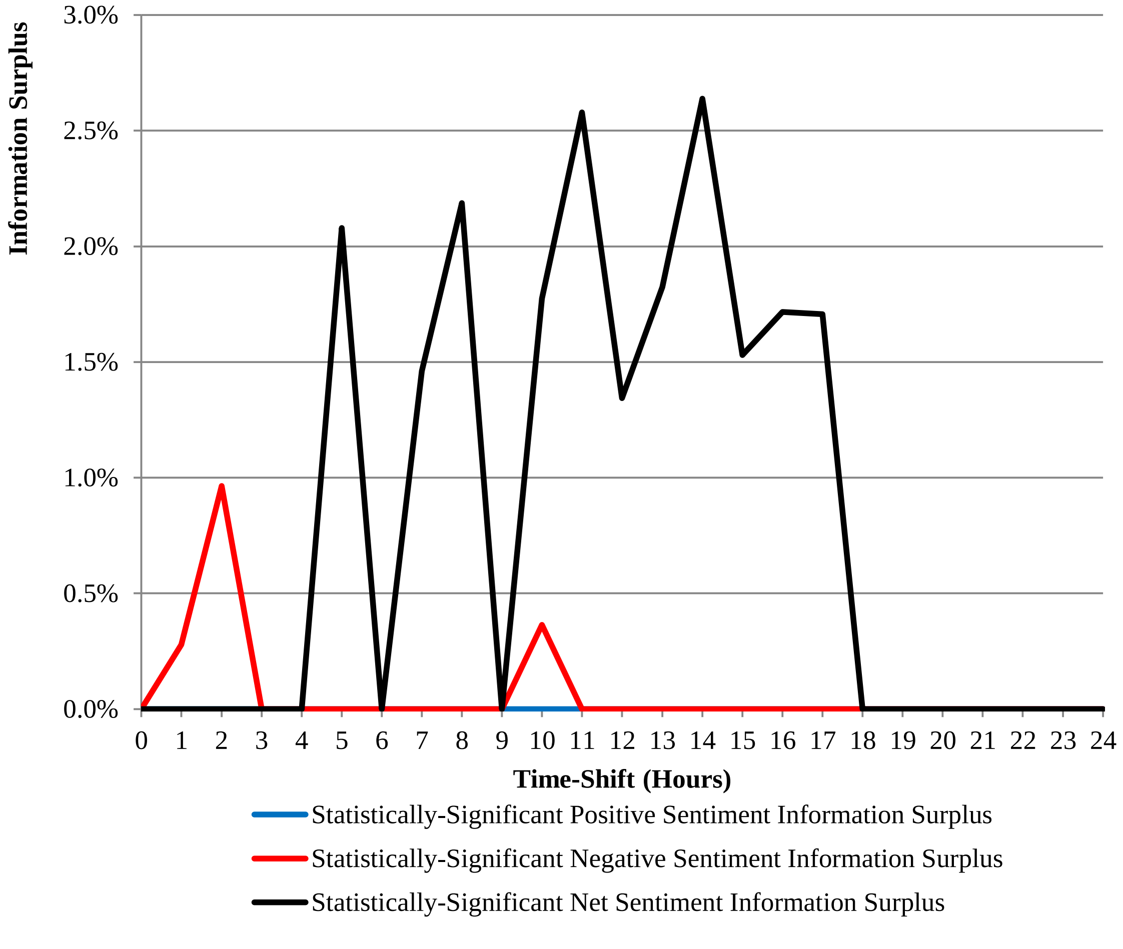


FIGURE 10: EXAMPLE SHOWING WHEN HOURLY CHANGES IN TWEET SENTIMENT DATA CAN LEAD A SECURITY'S HOURLY RETURNS IN A STATISTICALLY-SIGNIFICANT MANNER

By way of example, Figure 10 demonstrates the statistically-significant leading information surplus between hourly changes in sentiment data for the Twitter Filter: “$AMZN” OR “Amazon” and the hourly returns of Amazon.com, Inc. CFDs. Here, the performances of the three different sentiment types (positive, negative and net) are shown, as produced by the SentiStrength[50] classifier. Instances where the information surplus is non-zero denotes: a leading time-shift for which hourly changes in the sentiment data contain more information about the asset’s returns ahead of time than at zero time-shift in a statistically-significant manner and also that these changes in the sentiment data are more leading than trailing. In such instances therefore the social media data does indeed precede the financial data.

Note that Figure 10 is not showing the actual mutual information for each time-shift – rather it is showing the information surplus values: the percentage increase in the information surplus for time-shifts >0 hours, relative to the information surplus for a time-shift of zero hours. When the information surplus is zero for a particular time-shift, this denotes that the mutual information between the social media data and the financial data is not statistically-significant, and thus shows that the social media data does not lead the financial data. Therefore, as with the case in Figure 10, it is possible for the net sentiment’s information surplus to be statistically-significant at a particular time-shift, whilst the positive and/or negative sentiments are not statistically-significant at the same time-shift. For example, consider the time-shift of 10-hours, at which point the information surplus values for the net sentiment and the negative sentiment are non-zero (and therefore statistically-significant), whilst the information surplus for the positive sentiment is statistically-insignificant, and therefore shown as zero. For this time-shift, the social media’s net and negative sentiments lead the financial data in a statistically-significant manner, whilst the positive sentiments do not.

5.5.4 Determining if social media message sentiments carry greater abilities to lead securities’ returns than social media message volumes

This study is concerned with evaluating whether hourly changes in sentiment data carry a greater ability to lead securities’ hourly returns than just hourly changes in Tweet volumes. Thus, for each financial-instrument/Twitter-Filter combination, experiments are first performed by considering changes in Tweet sentiments as evaluated against changes in assets’ prices: $\Delta_{\mathrm{sentiment}}$ vs. the Δ_{price}.

For each financial-instrument/Twitter-Filter combination, the experiments are then repeated by considering changes in Tweet message volumes, as evaluated against changes in assets' prices: $\Delta_{\text{message volume}}$ vs. the Δ_{price}.

In addition, to echo past studies which evaluate Google Search Trends[13-15] and Yahoo! search engine data[16] message volumes against financial market performance, the experiments are also repeated to consider the relationships between changes in Tweet message volumes and assets' absolute returns. Thus, for each financial-instrument/Twitter-Filter combination, the study also considers changes in Tweet message volumes as evaluated against absolute changes in the asset's prices: $\Delta_{\text{message volume}}$ vs. the $|\Delta_{\text{price}}|$.

These experiments allow for the identification of the extent to which changes in Twitter message sentiments can lead securities' returns over and above what is attainable by the evaluation of changes in Twitter message volumes.

5.6 Functions of the study's software programs

A series of MATLAB-based analysis frameworks were designed for use in this study, to analyse the data produced by the TCF. An illustration of these software packages' interactions is presented in Figure 11, and features discussed subsequently.

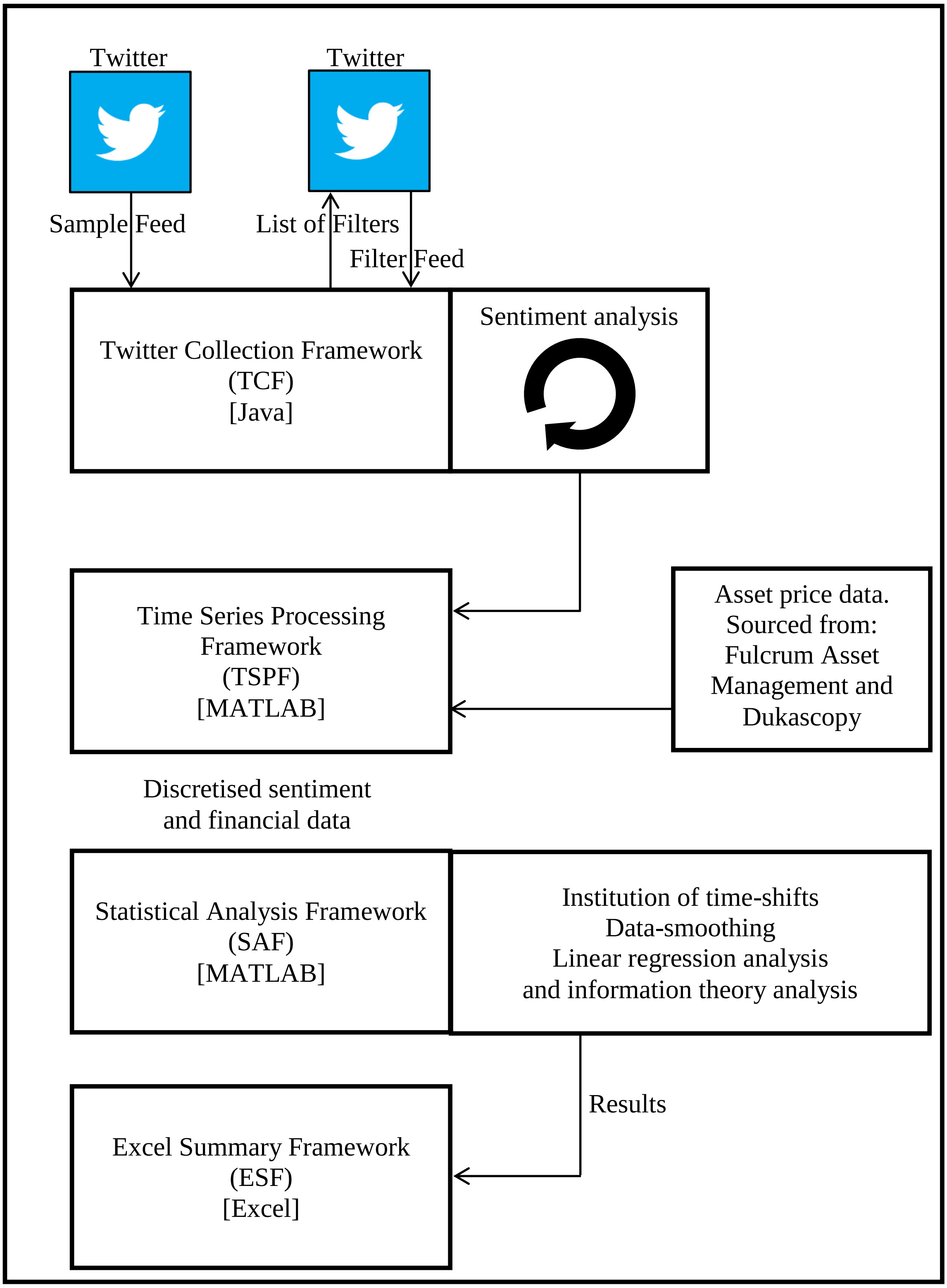
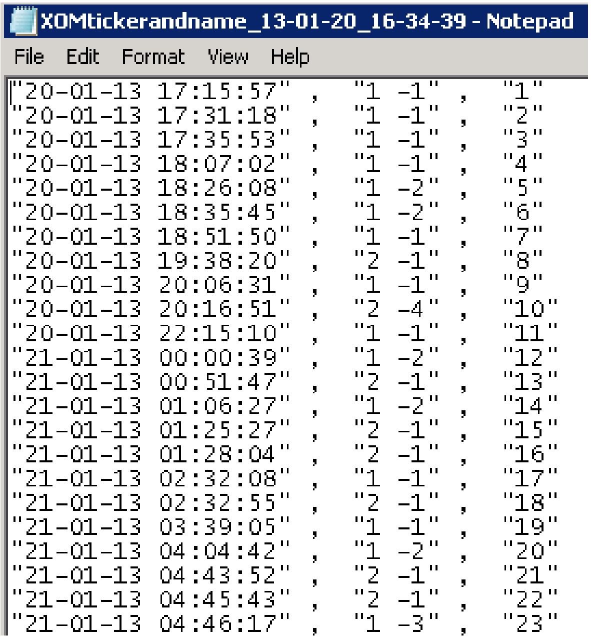


FIGURE 11: INTERACTIONS OF THE SOFTWARE FRAMEWORKS DEVELOPED FOR THE STUDY

As discussed in Chapter 4.2 and Chapter 4.3, and listed in Table 1, a series of string-based and geographical Twitter filters were set-up to collect and filter relevant messages from Twitter's Gardenhose Feed. The data were collected over a 3-month period from 11th December 2012 to 12th March 2013, as explained in Chapter 5.2. An example of the data produced by the Twitter Collection Framework (TCF) is shown below:



```
"20-01-13 17:15:57" , "1 -1" , "1"
"20-01-13 17:31:18" , "1 -1" , "2"
"20-01-13 17:35:53" , "1 -1" , "3"
"20-01-13 18:07:02" , "1 -1" , "4"
"20-01-13 18:26:08" , "1 -2" , "5"
"20-01-13 18:35:45" , "1 -2" , "6"
"20-01-13 18:51:50" , "1 -1" , "7"
"20-01-13 19:38:20" , "2 -1" , "8"
"20-01-13 20:06:31" , "1 -1" , "9"
"20-01-13 20:16:51" , "2 -4" , "10"
"20-01-13 22:15:10" , "1 -1" , "11"
"21-01-13 00:00:39" , "1 -2" , "12"
"21-01-13 00:51:47" , "2 -1" , "13"
"21-01-13 01:06:27" , "1 -2" , "14"
"21-01-13 01:25:27" , "2 -1" , "15"
"21-01-13 01:28:04" , "2 -1" , "16"
"21-01-13 02:32:08" , "1 -1" , "17"
"21-01-13 02:32:55" , "2 -1" , "18"
"21-01-13 03:39:05" , "1 -1" , "19"
"21-01-13 04:04:42" , "1 -2" , "20"
"21-01-13 04:43:52" , "2 -1" , "21"
"21-01-13 04:45:43" , "2 -1" , "22"
"21-01-13 04:46:17" , "1 -3" , "23"
```

FIGURE 12: EXAMPLE OF THE RAW DATA PRODUCED BY THE TWITTER COLLECTION FRAMEWORK (TCF)

Figure 12 shows an excerpt from an output file produced by the TCF (see Chapter 4.2), showing the resultant sentiment scores produced for Tweets matching a particular Twitter-Filter, which in this example is "$XOM" (the industry ticker-ID for Exxon Mobil). This figure shows that the data at this stage consists of time-stamps of non-discretised sentiments.

There are four columns of data per row: non-discretised continuous timestamp; positive and negative sentiments[a]; and row number. Rows are listed in ascending chronological order.

[a] The sentiment classifier used in the study, SentiStrength, produces two classifications per text string: a positive component, and a negative component. Further descriptions of the functionality of the SentiStrength classifier are available in Chapter 5.1.

The following is an example of the raw Tweet from Figure 12:

"Exxon Mobil disappoints, shares down 3.6% premarket. $XOM"

Using SentiStrength[50] the classifier employed in this study produces the following rankings for the string above:

- A positive sentiment score of **+1**, which is ranked by the classifier on a scale from +1 (least positive) to +5 (most positive);
- A negative sentiment score of **-3**, which is ranked by the classifier on a scale of -1 (least negative) to -5 (most negative);
- A net sentiment score of **-2**, which is produced by the summation of the negative and positive sentiment scores and is therefore ranked on a scale of -4 (most negative) to +4 (most positive).

5.6.1 Time Series Processing Framework (TSPF)

The Time Series Processing Framework (TSPF) has the following user-controlled options:

1. Social media data read-in selection. The user selects if he wishes to read in raw social media data for a particular financial-instrument/Twitter-Filter combination from the Twitter Collection Framework (TCF) for the first time, or if he wishes to open data from the TCF that has been read-in on a previous occasion. Reading data for the first time is more time-consuming as the TSPF has to convert .txt file data into MATLAB's own .m file data line by line, and this takes place at a rate of up to 2,500 rows per second on a standard desktop machine. Opening the pre-read data any subsequent time is near-instantaneous. There is no limitation on the size of the data files which can be read-in.
2. Financial data read-in selection. The user selects the underlying file which contains the raw price data for a particular financial-instrument/Twitter-Filter combination. Financial data are sourced either from Dukascopy (in which case the data are in the form of a CSV), or from Fulcrum Asset Management (in

which case the data are the form of an .m file), as discussed in Chapter 4.3. Whilst there is no restriction on the granularity of the financial data that can be used, all financial data considered in this study were presented in 5-minute tick intervals.

3. Discretisation-window selection. The user selects the size of the window into which the social media and financial data are aggregated. This allows for the conversion of raw data, which is continuous, into discretised time frames, as discussed in Chapter 5.3. The choice of discretisation frequency in the financial services industry is often ad-hoc, typically dictated by the observation intervals of the available data[79]. As discussed in Chapter 4.1, the development of SocialSTORM[57] provided preliminary access to Twitter data for initial exploration of the relationships between social media data and financial data. Whilst the Twitter data provided by SocialSTORM which was continuous, as is the case with the TCF, the financial data used during this preliminary investigation was not available to discretised resolutions smaller than an hour[a80]. Based on this past data limitation, it was decided that relationships between Twitter data and financial data would be evaluated as discretised to the hourly level, followed by testing the robustness of the relationships at different discretisation levels (as discussed in Chapter 7.1).

   For example, if the user selects the window to be 1-hour in size, the system performed the following calculations:

   a) A discretised time-series T of time-stamps with elements $\mathrm{T_i}$ is created, where $\mathrm{T_1} =$ 00:00:00 on 11th December 2012 and $\mathrm{T_n} =$ 23:59:59 on 11th March 2013 (bringing the data-capture period up to 12th March 2013, giving a total of 90 days).

   b) The number of periods per 24-hours is determined as a function of the desired window size, W when expressed in hours (in this example, 1):

$$\mathrm{N_{periods}} = \frac{24}{1}$$

   The number of elements in the discretised time-series T is therefore:

[a] Financial data used for the preliminary investigation was sourced from Thomson Reuters and from Fulcrum Asset Management, and was discretised to hourly windows due to the unavailability of higher-resolution data.

$$\mathrm{T_n} = \mathrm{N_{periods}} \times 90 = 24 \times 90 = 2160$$

c) It is then identified whether the input data time-series of price, sentiment and message volume, $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$ belong to each location in the discretised time-series T. An input data-point I is deemed to belong to a location in the discretised time-series T if its time-stamp is between up to and including the time-stamp for the current location in the discretised time-series, $\mathrm{T_i}$, and above but not including the time-stamp for the chronologically previous location in the discretised time-series, i.e., $\mathrm{T_{i-1}}$.

d) For each location in the discretised time-series T, the discretised means of the values for each of the corresponding input data series of price, sentiment and message volume, $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$ are determined. Denoted $\overline{\mathrm{D_{price_{T_n}}}}$, $\overline{\mathrm{D_{sentiment_{T_n}}}}$ and $\overline{\mathrm{D_{message\ volume_{T_n}}}}$ respectively, these are calculated as:

$$\overline{\mathrm{D_{price_{T_i}}}} = \frac{\mathrm{I_{price_1}} + \mathrm{I_{price_2}} + \cdots \mathrm{I_{price_n}}}{\mathrm{n}}$$

$$\overline{\mathrm{D_{sentiment_{T_i}}}} = \frac{\mathrm{I_{sentiment_1}} + \mathrm{I_{sentiment_2}} + \cdots \mathrm{I_{sentiment_n}}}{\mathrm{n}}$$

$$\overline{\mathrm{D_{message\ volume_{T_i}}}} = \frac{\mathrm{I_{message\ volume_1}} + \mathrm{I_{message\ volume_2}} + \cdots \mathrm{I_{message\ volume_n}}}{\mathrm{n}}$$

e) Finally, the changes in these discretised mean values of $\mathrm{I_{price}}$, $\mathrm{I_{sentiment}}$, $\mathrm{I_{message\ volume}}$ are then calculated. Denoted $\Delta\overline{\mathrm{D_{price_{T_n}}}}$, $\Delta\overline{\mathrm{D_{sentiment_{T_n}}}}$ and $\Delta\overline{\mathrm{D_{message\ volume_{T_n}}}}$ respectively, these are calculated as

$$\Delta\overline{\mathrm{D_{price_{T_i}}}} = \overline{\mathrm{D_{price_{T_i}}}} - \overline{\mathrm{D_{price_{T_{i-1}}}}}$$

$$\Delta\overline{\mathrm{D_{sentiment_{T_i}}}} = \overline{\mathrm{D_{sentiment_{T_i}}}} - \overline{\mathrm{D_{sentiment_{T_{i-1}}}}}$$

$$\Delta\overline{\mathrm{D_{message\ volume_{T_i}}}} = \overline{\mathrm{D_{message\ volume_{T_i}}}} - \overline{\mathrm{D_{message\ volume_{T_{i-1}}}}}$$

In this manner, this methodology not only discretises the input data, but also normalises the data by the volume of data-points for each element in the time-series T.

f) Note, the values of $\Delta\overline{\mathrm{D}_{\mathrm{price}_{\mathrm{T}_1}}}$, $\Delta\overline{\mathrm{D}_{\mathrm{sentiment}_{\mathrm{T}_1}}}$ and $\Delta\overline{\mathrm{D}_{\mathrm{message\ volume}_{\mathrm{T}_1}}}$ (i.e., for element T_1) are empty as these are the first entries in the discretised time-series T and therefore there are no prior elements from which to calculate the changes in these discretised mean values of $\mathrm{I}_{\mathrm{price}}$, $\mathrm{I}_{\mathrm{sentiment}}$, $\mathrm{I}_{\mathrm{message\ volume}}$.

The TSPF also calculates the net sentiment for each Tweet, as described in Chapter 5.1. This is calculated by subtracting the negative sentiment from the positive sentiment for each message, and is ranked on a scale of -4 (most negative) through 0 (neutral) to +4 (most positive).

A full copy of the code underpinning the TSPF is available in the Appendix (see Chapter 11.2).

5.6.2 Statistical Analysis Framework (SAF)

The Statistical Analysis Framework (SAF) was developed to read in data produced by the Time Series Processing Framework.

The time-stamped discretised social media data from the TSPF were found to contain repeating patterns within the data. A method for addressing these repeating patterns was implemented in the Statistical Analysis Framework (SAF), and is discussed in Chapter 5.6.2.1.

The SAF also instituted time-shifts between the social media and financial datasets, as discussed in Chapter 5.6.2.2.

Finally, the SAF performed both linear regression and information theory analyses the dataset as per the theories detailed in Chapter 5.3.1 and Chapter 5.3.2.

A full copy of the code underpinning the SAF is available in the Appendix (see Chapter 11.2).

5.6.2.1 Time-series decomposition

Consider that a time-series, or sequence of data ordered in time, can be described as a function of notional sub-components. The classical decomposition[98] of a time-series $\mathrm{Y_t}$ is:

$$\mathrm{Y_t = TC_t + S_t + R_t}$$

Where:

- $\mathrm{Y_t}$ is the time series value at period t;
- $\mathrm{TC_t}$ is the trend-cycle component of the data, which describes the underlying movement in a time-series. This can be positive, absent, or negative, and can be linear or non-linear[99]. The trend-cycle component represents the data this study is concerned with isolating;
- $\mathrm{S_t}$ is the seasonal effect within the data, which describes periodic fluctuations in the data attributed to seasonal factors such as: the quarter of the year[99] which can be influenced by annual economic and sociological-driven cycles;
- $\mathrm{R_t}$ is the random component of the data. It is the remainder of the time-series once the trend and cyclical components have been removed, typically treated as white noise[a][98, 100] in time-series decomposition.

Preliminary exploration of the hourly discretised data produced by the TSPF, an example of which is seen in Figure 13, showed that the social media data time-series contains a repeating cyclical variation component. Figure 13 shows an excerpt of the actual sentiments (**not** $\Delta_{\mathrm{sentiment}}$) of string-unfiltered Tweets from the US, aggregated over 1-hour discretisation windows:

[a] White noise is a random process within a time-series with zero autocorrelation. It is used as a common model of noise in time-series analysis.

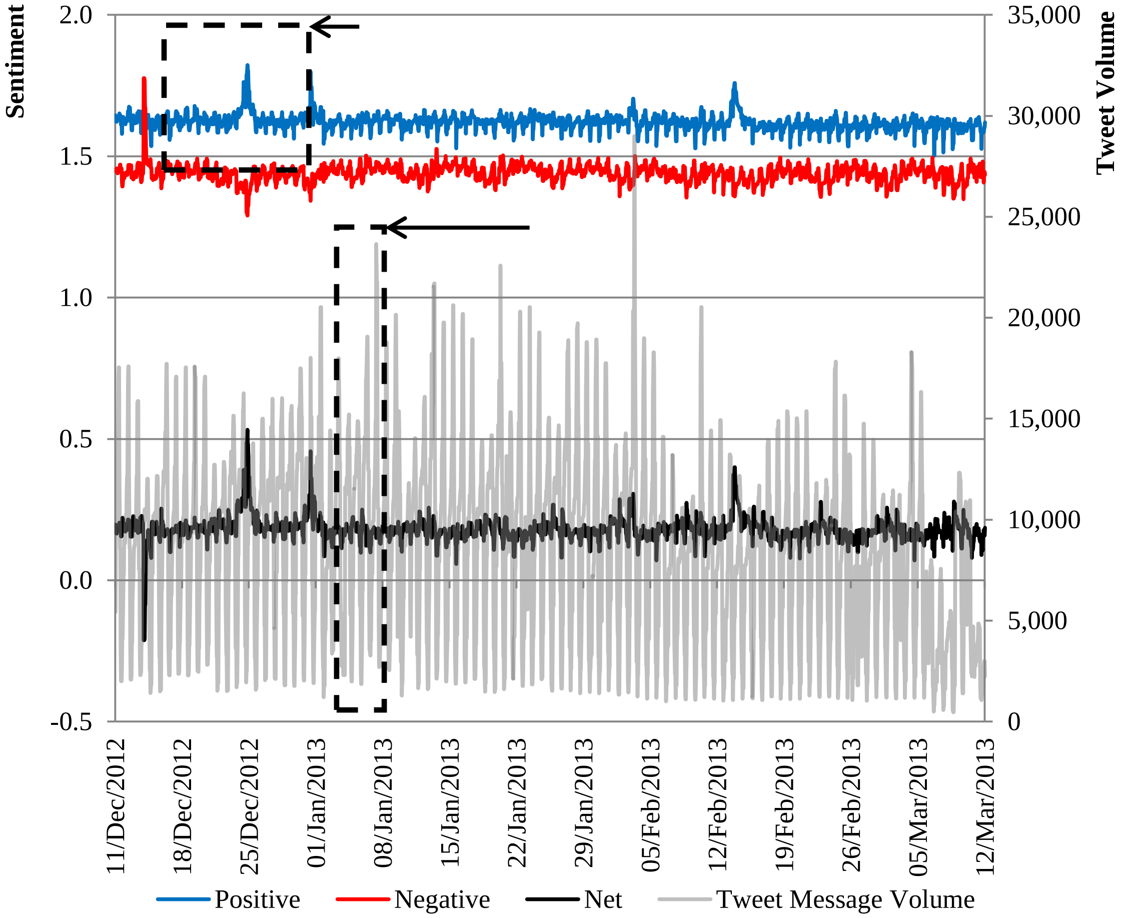


FIGURE 13: EXAMPLE OF SEASONALITY WITHIN SOCIAL MEDIA DATA FOR STRING-UNFILTERED TWEETS FROM THE US DISCRETISED TO HOURLY WINDOWS

Figure 13 shows **hourly** data, which demonstrates cyclical patterns in the both Twitter sentiments and Twitter volumes. The cyclical patterns of the Twitter message volumes can be explained by the fact that internet users are less likely to Tweet during night hours than during daylight hours. This is particularly evident in the data shown this figure since it is sourced from geographically-filtered (US) data, rather than string-filtered Twitter data. The existence of cyclical variation within Twitter message volumes further underpins the existence of cyclical variation within Twitter sentiments. For example, within Twitter's demographic (as discussed in Chapter 3.2), prevailing and repeating positive moods are most dominant during evening hours when individuals are likely to be socialising. Similarly, prevailing and repeating negative moods could be most dominant during morning hours when individuals are likely to be heading to work.

By assessing the components of the classical time-series decomposition formula $\mathrm{Y_t = TC_t + S_t + R_t}$, the cyclical variation in the social media data was assessed by the SAF before establishing its relationships with the financial data:

- $\mathrm{S_t}$: The seasonal effect within the data. As discussed in Chapter 5.2, the lack of availability of historic data from Twitter's network prevents the identification of historical seasonal effects within such time-series. This prevents the use of past seasonal effects for the removal of future seasonal effects within such time-series. Thus, the dataset considered by the study is intentionally one annual quarter – a length of time that captures a range of market & holiday conditions whilst not exceeding an annual quarter. Analysing a dataset of greater than an annual quarter in length would require access to historic data[a] from Twitter's network for use in mapping quarterly seasonality effects with the aim of removing them. Therefore, an assumption is made that the seasonal effect within the data cannot be addressed mathematically given the need to limit the data-collection period to avoid experiencing substantial changes to Twitter's product (see Chapter 5.2) – the seasonality effect is instead minimised by the selection of a purposefully-short chronological time-frame of data. Therefore the time-series decomposition formula is altered to $\mathrm{Y_t = TC_t + R_t}$.

- $\mathrm{R_t}$: The random component of the data. Since this is typically treated as white noise[98, 100] in time-series decomposition, it therefore has a zero-mean, and thus is equally likely to be positive or negative. By removing this component from the altered time-series decomposition formula $\mathrm{Y_t = TC_t + R_t}$, the trend-cycle component $\mathrm{TC_t}$ remains.

The random component of the data can be removed by averaging elements of the time-series. A classical method of time series decomposition is the use of moving-average smoothing[101], which is a method of arithmetic data averaging which drops chronologically older observations to include new observations. Averaging elements of the social media data provides a clearer view of the true underlying behaviour of the time-series by eliminating the randomness in the data and leaving a smoothed trend-cycle component.

[a] Such historic data was unavailable at the time of conducting the study's experimental work.

The commonly-employed Simple Moving Average method, which uses a moving window which encapsulates observations on either side of the data point in question, is given as:

$$\mathrm{MA}_{\mathrm{t}} = \frac{1}{\mathrm{k}} \sum_{-\mathrm{m}}^{\mathrm{j=m}} \mathrm{Y}_{\mathrm{t+j}}$$

Where:

- k is the order of the moving average;
- $\mathrm{m} = \frac{(\mathrm{k}-1)}{2}$.

This method however is not suited to this study as it will inherently consider social media data ahead of time: half of the elements in question are for time-periods prior to the element in question and the other half are for elements occurring for time-periods ahead of the element in question. Considering that this study seeks to identify instances in which social media data can lead financial data, one cannot base such analyses on the smoothing of future data. What is therefore needed is a method for smoothing which only considers past data.

The study therefore used a backwards-looking Simple Moving Average (SMA), and in this manner only considers elements in the time-series which have occurred in the past, thus preserving the integrity of using past social media data to assess if future financial data can be mapped.

The implementation of the backward-looking Simple Moving Average in the SAF required the identification of the number of elements for the window size: too many, and the data will be over-smoothed; too few, and the data will be under-smoothed. To address this issue, the autocorrelation technique is employed, which allows for estimation of the dominating frequency within the social media time-series[102].

Autocorrelation is a representation of the amount of similarity of an observation within a time-series, and another observation within the same time-series, as a function of time separation between such observations[103]. For a discrete process for which there are n observations $\{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}\}$, autocorrelation is obtained as:

$$\widehat{\mathrm{R}}(\mathrm{k}) = \frac{1}{(\mathrm{n}-\mathrm{k})\sigma^2}\sum_{\mathrm{t}=1}^{\mathrm{n}-\mathrm{k}}(\mathrm{X}_\mathrm{t}-\mu)(\mathrm{X}_{\mathrm{t}+\mathrm{k}}-\mu)$$

Where:

- μ is the arithmetic mean of the data;
- σ^2 is the variance of the data.

The estimation of the dominating frequency of a discrete signal can be performed by the identification of the largest peak in the autocorrelation function of a time-series occurring at a non-zero lag[104] – by definition, the signal is at its peak autocorrelation at a lag of zero[105]. Compared to the use of the Fourier transform[a][106], this methodology is more accurate since the resolution is not limited by the number of samples considered[107].

Therefore, to identify the largest peak in the autocorrelation of the social media time-series which occurs at a non-zero lag, Twitter sentiment data discretised to hourly windows is used, as collected by the methods described throughout this study. Consider the figure below:

[a] This is a mathematical transformation employed for conversion of signals between time domains and frequency domains.

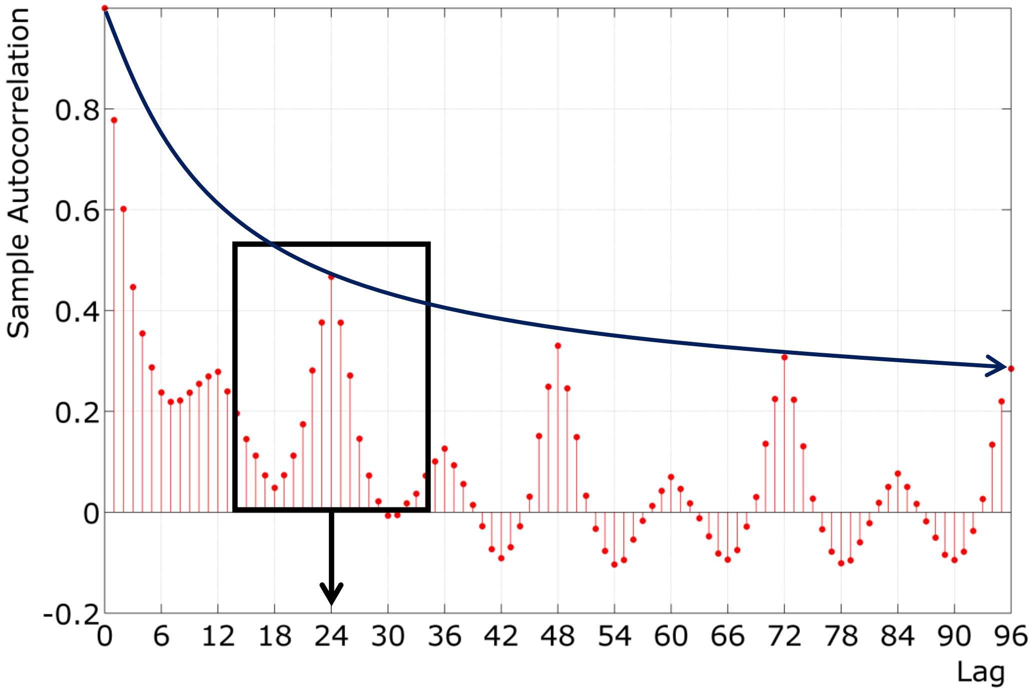


FIGURE 14: AUTOCORRELATION WITHIN STRING-UNFILTERED SENTIMENT DATA OF TWEETS FROM THE US DISCRETISED TO 1-HOUR WINDOWS

Figure 14 shows an autocorrelation plot for US string-unfiltered net sentiment data discretised over 1-hour periods. The black box shows the peak non-zero autocorrelation within the sentiment time-series (excluding the full autocorrelation occurring at a lag of zero[104]). The blue curve shows the decreasing autocorrelation peaks which occur at lags of 24-hour multiples. The largest autocorrelation peak takes place at a lag of 24 hours. This is a significant observation, demonstrating that Twitter sentiment data is autocorrelated at a lag of 24-hours – an observation that is seen throughout the study's 3-month dataset. This observation is used as a basis for the selection of the number of elements in the study's backward-looking Simple Moving Average (SMA) calculations.

Since the peak non-zero-lag autocorrelation takes place at a period of 24-hours, 24 elements are used for the size of the SMA window. The robustness of the study's results is considered in Chapter 7, and justification is given in Chapter 7.4 regarding the reason why the size of the SMA window cannot be relaxed.

The 24-hour a backward-looking Simple Moving Average (SMA) was implemented in the following manner:

- For each element in the social media time-series, the arithmetic mean of the preceding twenty-three data points and the element in question was calculated.

For example, for element 42 of the discretised social-media time-series D: $\mathrm{SMA}_{\mathrm{i}=42} = \frac{\mathrm{D}_{42}+\mathrm{D}_{41}+\cdots\mathrm{D}_{19}}{24}$.

- However, for the first twenty-three entries in the social media data time-series – for which there are less than twenty-four preceding elements – the SMA for each such entry is calculated based on the arithmetic mean of the element itself and all available chronologically-preceding elements, up until the first in the time-series. For example, for element 13 of the social-media time-series series D: $\mathrm{SMA}_{\mathrm{i}=13} = \frac{\mathrm{D}_{13}+\mathrm{D}_{12}+\cdots\mathrm{D}_{1}}{13}$.

The effect of instituting a 24-hour backwards looking Simple Moving Average on the sentiment data underpinning Figure 13 (US string-unfiltered sentiment data discretised over 1-hour periods) is shown below in Figure 15:

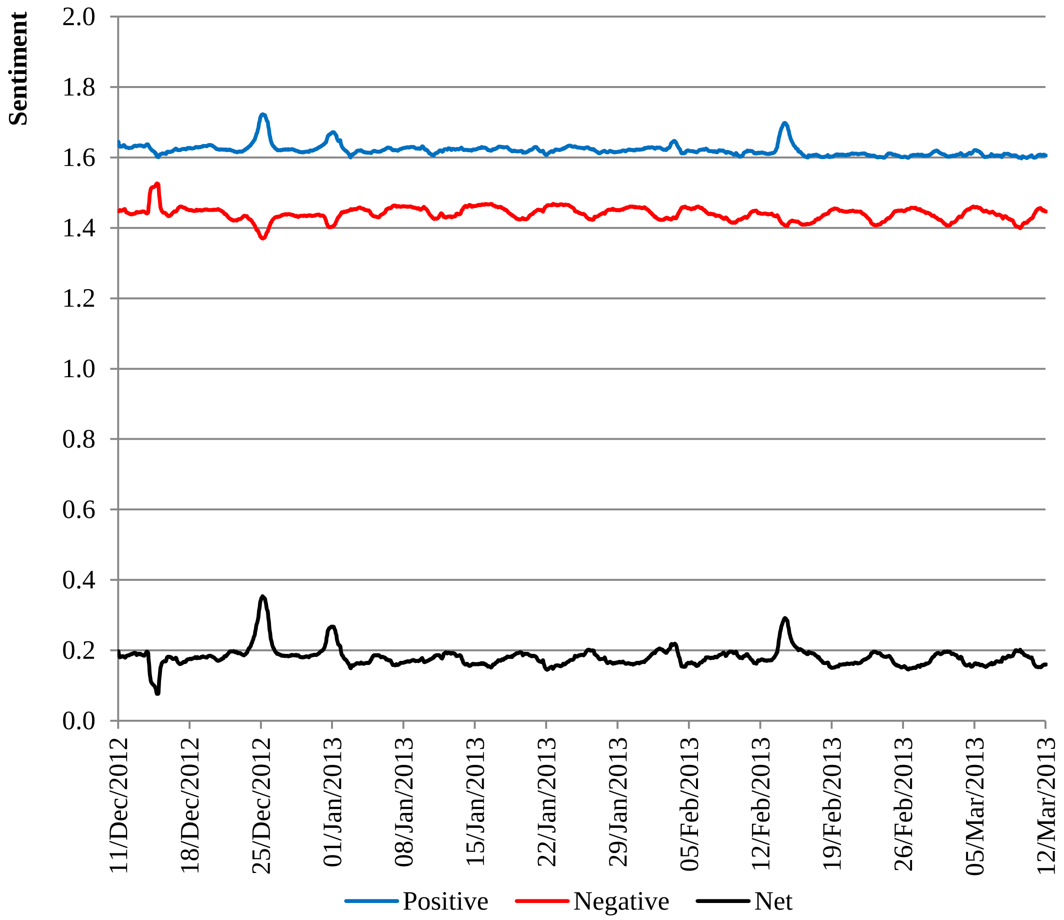


FIGURE 15: EFFECT OF 24-HOUR BACKWARD-LOOKING MOVING AVERAGE SMOOTHING ON US-SOURCED STRING-UNFILTERED TWITTER MESSAGE SENTIMENTS

The implementation of backwards-looking Simple Moving Average windows to the social media data in the SAF allowed for the isolation of the TC_t trend-cycle component. An example of this resultant isolation is shown in Figure 15. It is this component of the social media data which was further evaluated against the financial data using linear regression analysis (Chapter 5.3.1) and information theory (Chapter 5.3.2) to assess if it is able to lead the latter.

5.6.2.2 Implementation of time-shifts

As discussed in Chapter 5.3.2, time-shifts were instituted between the social media data and the financial data. A time-shift is an artificial chronological offset between the two time-series used for the calculation of the extent to which social media data is able to lead financial data.

Given that the fundamental discretisation window in this study was 1-hour (as discussed in Chapter 5.3), time-shifts were instituted in hourly multiples. The maximum time-shift considered in this study is +/– 24-hours[a], on the hour. The robustness of the study's results was tested against the effect of offsetting the hourly time-shift window such that it is not centred on the hour. This was achieved as a by-product of experiments to explore the effect of the aggregation the study's dataset to discretisation windows not centred on the hour, but rather off the hour (see Chapter 7.2).

The time-shifts instituted between the social media data and the financial data were both positive (such that social media data leads financial data) and negative (such that social media data trails financial data).

[a] The scope of the study is the exploration of social media's ability to lead financial data up to 24-hours ahead of time.

5.6.3 Excel Summary Framework (ESF)

The Excel Summary Framework is a collection of interlinked Microsoft Excel VBA[a] workbooks that aggregate results data from the SAF based on the analyses as set out in Chapter 5.3.1 and Chapter 5.3.2. It amalgamated and condensed the results produced by the SAF into coherent summaries through the use of automated VBA scripts.

These amalgamated results could then be visualised by the ESF – an example of the visualisation is shown in Figure 16. Here, the sentiment on the company Apple, Inc. as produced by the TCF (see Chapter 4.2) using the string-based Twitter Filter: “Apple” AND/OR “$AAPL”, is visualised in conjunction with Tweet message volumes. In this example, these sentiments and Tweet message volumes were discretised hourly, but no data-smoothing backward-looking simple moving average (SMA) was applied (see Chapter 5.6.2 for an explanation of time series decomposition and the necessity for data-smoothing). Figure 16 also shows the price of Apple, Inc. CFDs over the same period.

Figure 17 shows data for the same financial-instrument/Twitter-Filter combination as for Figure 16, with a 24-hour backward-looking rolling simple moving average (SMA) applied to the social media data as per the methodology described in Chapter 5.6.2.

Figure 18 shows hourly changes in the data on Apple, Inc., i.e., the $\Delta_{\text{sentiment}}$, the $\Delta_{\text{message volume}}$ and the Δ_{price} (i.e., returns) of Apple, Inc., CFDs.

[a] VBA: Visual Basic for Applications. This is an event-driven programming language based on Visual Basic, and can be used by Microsoft Office applications to create custom automated macros.

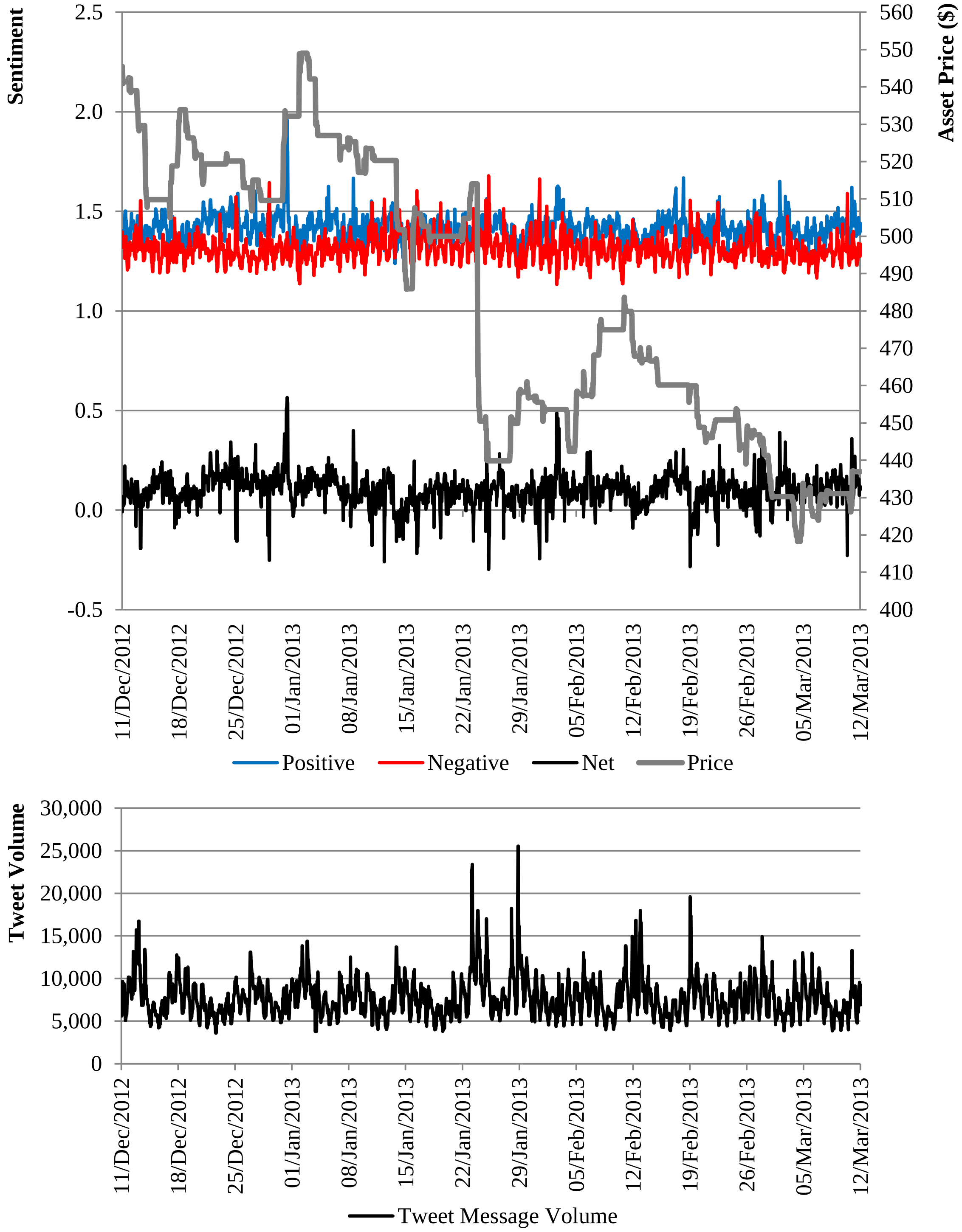


Figure 16: Tweet message sentiments and volumes vs. CFD prices for Apple, Inc., discretised hourly, with no data-smoothing applied to the social media time-series

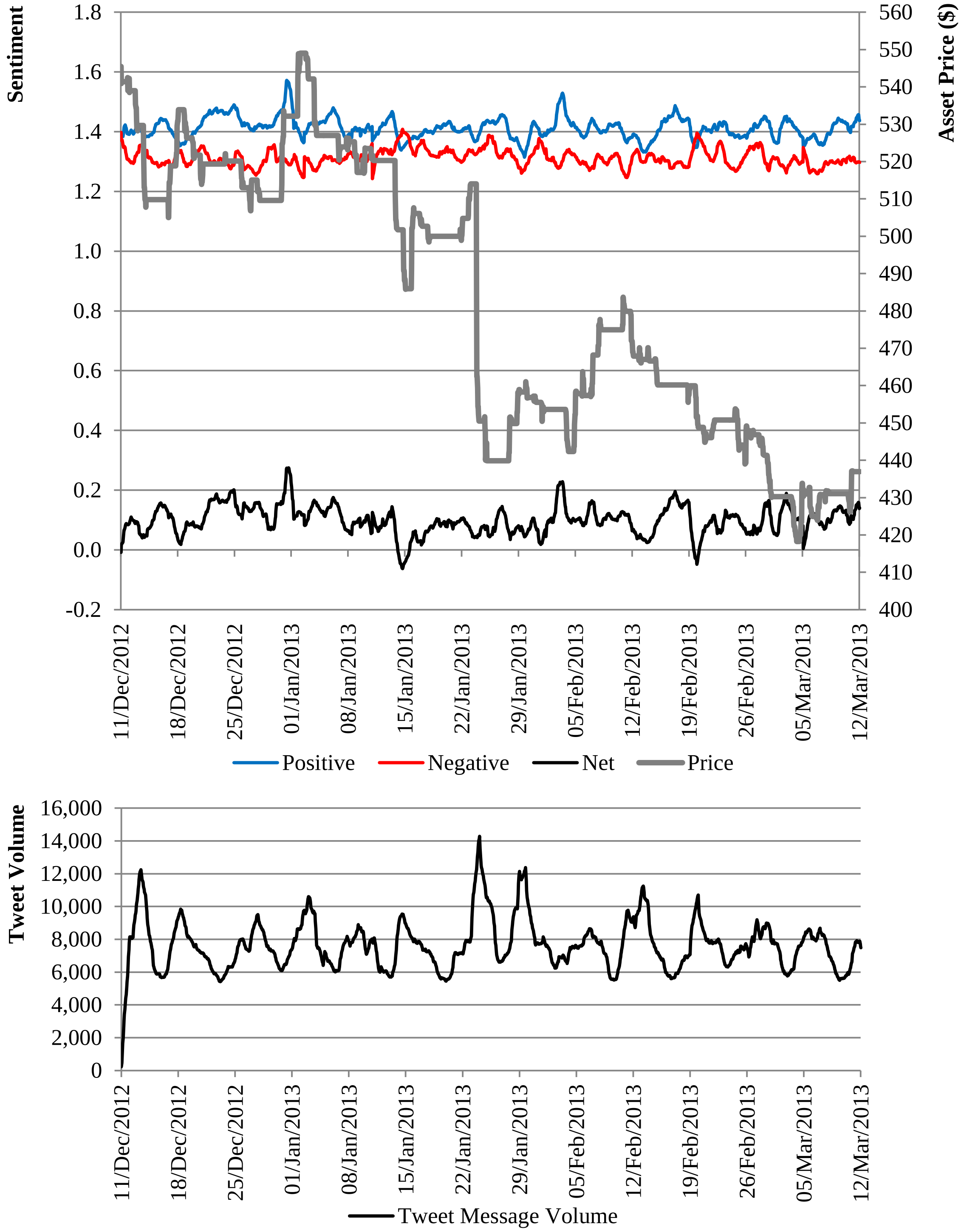


Figure 17: Tweet message sentiments and volumes vs. CFD prices for Apple, Inc., discretised hourly with a 24-hour backward-looking simple moving average (SMA) applied to the social media time-series

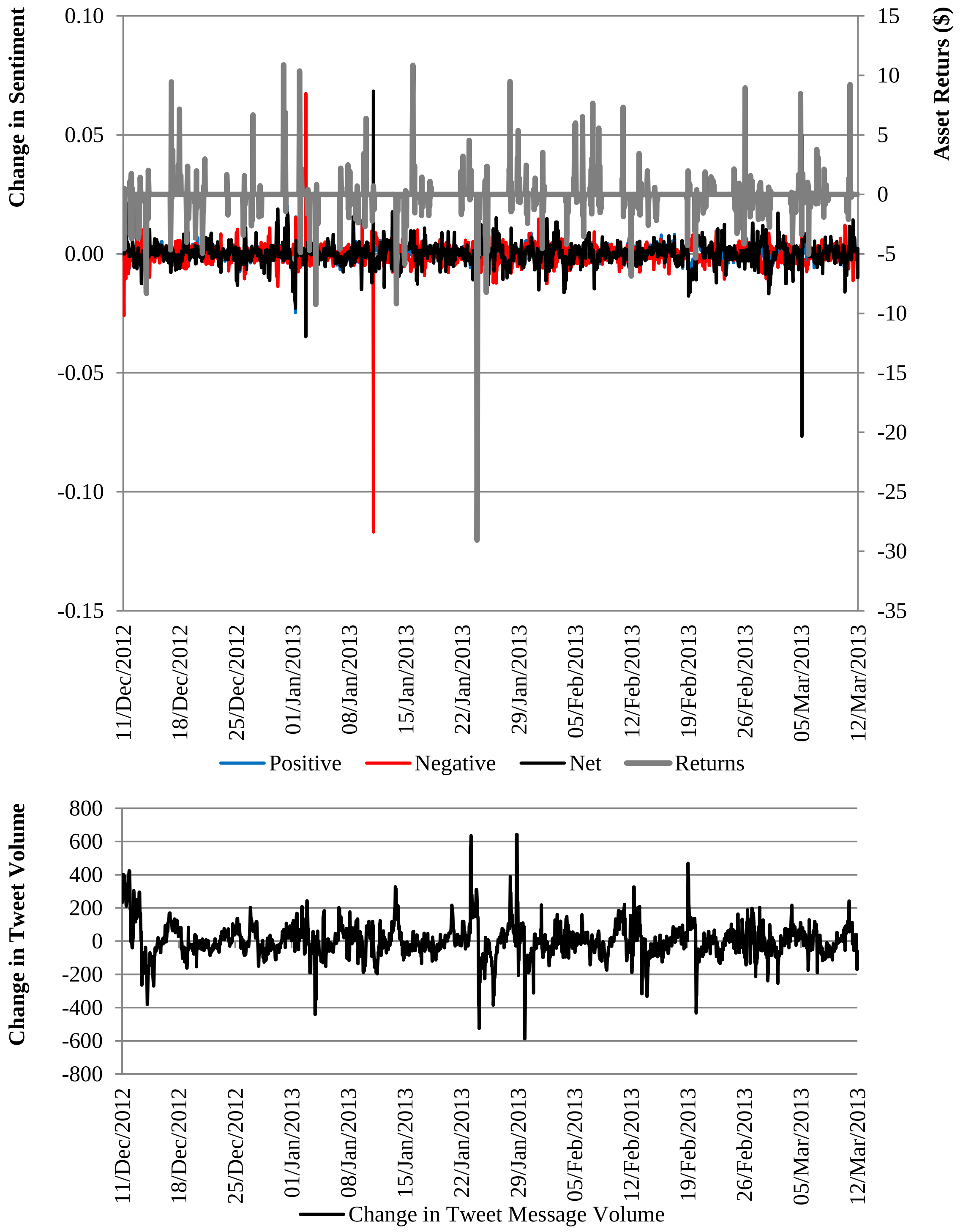


Figure 18: Hourly changes in Twitter message sentiments vs. hourly CFD returns for Apple, Inc., discretised hourly, with a 24-hour backward-looking simple moving average (SMA) applied to the social media time-series

6 RESULTS

The results of the study are presented to quantify and demonstrate the extent to which social media data can lead financial data.

6.1 Excluding financial-instrument/Twitter-Filter combinations based on impractical message volumes

Practical considerations for the prospect of trading from Twitter data have to be acknowledged – market insight from Tweets is only valuable if it can be applied practically.

Twitter's network operates using strict protocol for data requests, imposing limits on the number of times access to its resources can be requested per time period. As detailed in Twitter's development documentation[a], the strictest such limit is 15 API requests per 15-minute period, i.e., a mean of 1 request per minute.

Therefore, should one wish to make practical use of Twitter data, i.e., trade from it, what would be needed is a trading-model which can maximise its ability to react to changes in the Tweets fed through the company's network by altering the data it requests from Twitter, without violating the firm's connection protocols.

Therefore, based on this strictest limit of 1 API request per minute – which translates to 1 change per minute to the data requested by a trading-model – the minimum average Tweet rate needed to utilise this limit is 1 Tweet per minute. Any less than 1 message per minute, and the trading-model would not be making full use of Twitter's rate limit of 15 API requests per 15-minutes.

This study therefore uses a minimum viable mean message volume of 1 Tweet per Twitter filter per minute over the investigation's 3-month dataset. Therefore, any financial-instrument/Twitter-Filter combinations which attract a mean message volume of less than 1 Tweet per minute are excluded by the study. Based on this message-volume filter, twenty-three of the forty-four financial-instrument/Twitter-Filter combinations are excluded.

[a] https://dev.twitter.com/docs/rate-limiting/1.1/limits

Another practical consideration is the use of Twitter filters which yield messages relating to the companies in question. This study therefore also excludes Twitter filters which reference companies whose names are only two characters in length, since these were found to attract messages not related to the companies in question.

Specifically, Tweets on the company 3M, Co. cannot be filtered accurately since the term “3M” attracts a large volume of messages that have no association with the firm. Similarly, the term “GE” – an often-used trading name of General Electric, Co. – attracts large volumes of messages that do not refer this firm either. These two financial-instrument/Twitter-Filters are therefore also excluded, giving a total of twenty-five excluded Twitter filters. Table 5 lists the study’s forty-four financial-instrument/Twitter-Filter combinations, and indicates whether they are included or excluded on the basis of attracting correct messages and the minimum viable mean message volume of 1 Tweet per Twitter filter per minute.

Filter ID	Instrument	Filter type	Mean minutely message volume	Permitted or excluded?
1	Apple, Inc. CFDs	Ticker-ID AND/OR Company Name	126.7	Permitted
2	Apple, Inc. CFDs	Ticker-ID	1.8	Permitted
3	Amazon.com, Inc. CFDs	Ticker-ID AND/OR Company Name	123.1	Permitted
4	Amazon.com, Inc. CFDs	Ticker-ID	0.3	Excluded
5	American Express, Co.	Ticker-ID AND/OR Company Name	0.9	Excluded
6	Bank of America, Corp.	Ticker-ID AND/OR Company Name	1.6	Permitted
7	Bank of America, Corp.	Ticker-ID	0.2	Excluded
8	Cisco Systems, Inc. CFDs	Ticker-ID AND/OR Company Name	4.0	Permitted
9	EURUSD CFDs	Ticker-ID	0.8	Excluded
10	EURUSD Futures	Ticker-ID	0.8	Excluded
11	GBPUSD CFDs	Ticker-ID	0.3	Excluded
12	GBPUSD Futures	Ticker-ID	0.3	Excluded
13	General Electric, Co.	Ticker-ID AND/OR Company Name	74.8	Unfilterable, therefore excluded*
14	General Electric, Co.	Ticker-ID	0.1	Excluded
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184.0	Permitted
16	Google, Inc. CFDs	Ticker-ID	0.5	Excluded
17	The Home Depot, Inc.	Ticker-ID AND/OR Company Name	1.9	Permitted
18	Hewlett Packard, Co.	Ticker-ID AND/OR Company Name	0.8	Excluded
19	Hewlett Packard, Co.	Ticker-ID	0.2	Excluded
20	IBM, Corp. CFDs	Ticker-ID AND/OR Company Name	5.8	Permitted
21	IBM, Corp. CFDs	Ticker-ID	0.1	Excluded
22	Intel, Corp. CFDs	Ticker-ID AND/OR Company Name	12.9	Permitted
23	Intel, Corp. CFDs	Ticker-ID	0.1	Excluded
24	Johnson & Johnson, Co.	Ticker-ID AND/OR Company Name	0.1	Excluded
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	Permitted
26	J.P. Morgan, Inc. CFDs	Ticker-ID	0.1	Excluded
27	Coca-Cola, Co. CFDs	Ticker-ID AND/OR Company Name	24.8	Permitted
28	Coca-Cola, Co. CFDs	Ticker-ID	0.0	Excluded
29	McDonald’s, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	Permitted
30	McDonald’s, Corp. CFDs	Ticker-ID	0.1	Excluded
31	3M, Co. CFDs	Ticker-ID AND/OR Company Name	9.4	Unfilterable, therefore excluded*
32	Microsoft, Corp. CFDs	Ticker-ID AND/OR Company Name	30.0	Permitted
33	Microsoft, Corp. CFDs	Ticker-ID	0.2	Excluded
34	Oracle, Corp. CFDs	Ticker-ID AND/OR Company Name	5.0	Permitted
35	Oracle, Corp. CFDs	Ticker-ID	0.0	Excluded
36	FTSE100 Index CFDs	UK Geographical	35.5	Permitted
37	FTSE100 Index Futures	UK Geographical	35.5	Permitted
38	S&P500 Index CFDs	US Geographical	142.7	Permitted
39	S&P500 Index Futures	US Geographical	142.7	Permitted
40	AT&T, Inc. CFDs	Ticker-ID AND/OR Company Name	0.7	Excluded
41	AT&T, Inc. CFDs	Ticker-ID	0.0	Excluded
42	Wal-Mart, Inc. CFDs	Ticker-ID AND/OR Company Name	5.5	Permitted
43	Exxon Mobil, Corp. CFDs	Ticker-ID AND/OR Company Name	0.2	Excluded
44	Exxon Mobil, Corp. CFDs	Ticker-ID	0.1	Excluded

TABLE 5: DO THE STUDY'S TWITTER FILTERS ATTRACT CORRECT MESSAGES AND SUFFICIENT MESSAGE VOLUMES?

6.2 Linear regression analysis results

The linear regression analysis methodology, as described in Chapters 5.3.1 and 5.4 was used as a measure of dependency to explore the extent to which social media data can lead the financial data. As discussed in Chapter 5.4, to assess the efficacy of using linear regression analysis as a measure of dependency for the study's dataset, a straightforward implementation of the measure was needed. The linear regression analysis experiments were therefore performed using no time-shift and hourly discretisation[a] of the data (with 24-hour backwards-looking SMA smoothing: see Chapter 5.6.2.1) to assess if there are any correlations between hourly changes in the datasets at no chronological lag. The results of this configuration were used to direct further study.

The results of these experiments are presented below for the 19 financial-instrument/Twitter-Filter combinations which attracted correct message volumes (as discussed in Chapter 6.1).

[a] See Chapter 5.3 for an explanation of why hourly discretisation windows were used.

6.2.1 Experiment 1: No binning

The results of the experiments using no binning, as described in Chapter 5.4 are presented below in Table 6. Here, Pearson's r correlations are shown between hourly changes in the positive, negative and net sentiments of Tweets using hourly discretisation and 24-hour backward looking SMA smoothing and hourly returns of the financial assets.

			Pearson's r correlation for:		
Filter ID	**Instrument**	**Filter**	**Positive sentiment**	**Negative sentiment**	**Net sentiment**
1	Apple, Inc. CFDs	$AAPL AND/OR “Apple”	-0.052	-0.100	0.021
2	Apple, Inc. CFDs	$AAPL	-0.131	-0.059	-0.052
3	Amazon.com, Inc. CFDs	$AMZN AND/OR “Amazon”	0.047	-0.033	0.062
6	Bank of America, Corp. CFDs	$BAC AND/OR “Bank of America”	-0.023	-0.082	0.046
8	Cisco Systems, Inc. CFDs	$CSCO AND/OR “Cisco”	0.002	0.008	-0.019
15	Google, Inc. CFDs	$GOOG AND/OR “Google”	-0.150	-0.014	-0.106
17	The Home Depot, Inc. CFDs	$HD AND/OR “Home Depot”	-0.021	0.025	-0.029
20	IBM, Corp. CFDs	$IBM AND/OR “IBM”	-0.059	-0.100	0.054
22	Intel, Corp. CFDs	$INTC AND/OR “Intel”	-0.053	-0.071	0.014
25	J.P. Morgan, Inc. CFDs	$JPM AND/OR “JPMorgan” AND/OR “JP Morgan”	-0.133	-0.158	0.072
27	Coca-Cola, Co. CFDs	$KO AND/OR “Coca-Cola” AND/OR “Coca Cola”	-0.115	-0.083	-0.002
29	McDonald’s, Corp. CFDs	$MCD AND/OR “McDonald’s” AND/OR “McDonalds”	-0.050	-0.125	0.073
32	Microsoft, Corp. CFDs	$MSFT AND/OR “Microsoft”	-0.039	-0.193	0.081
34	Oracle, Corp. CFDs	$ORCL AND/OR “Oracle”	-0.005	0.046	-0.046
36	FTSE100 Index CFDs	String-unfiltered UK Tweets	-0.062	0.106	-0.098
37	FTSE100 Index Futures	String-unfiltered UK Tweets	-0.154	0.043	-0.125
38	S&P500 Index CFDs	String-unfiltered US Tweets	0.098	0.003	0.058
39	S&P500 Index Futures	String-unfiltered US Tweets	0.002	0.040	0.001
42	Wal-Mart, Inc. CFDs	$WMT AND/OR “Wal-Mart” AND/OR “Wal Mart”	-0.003	0.151	-0.187

TABLE 6: RESULTS OF LINEAR REGRESSION ANALYSIS WITH NO BINNING

Table 6 shows the Pearson's r correlations between Twitter sentiment data and assets' returns. For the no-binning configuration and for this experiment's parameters, the detected correlations show no or negligible relationships according to accepted interpretations of the values of Pearson's r[108,109].

6.2.2 Experiment 2: Binning by non-zero trading volume

The results of the experiments using binning by non-zero trading volume, as described in Chapter 5.4 are presented below in Table 7. Here, correlations are shown between hourly changes in the positive, negative and net sentiments of Tweets using hourly discretisation and 24-hour backward looking SMA smoothing and hourly returns of the financial assets.

The number of bins identifying chronological instances of non-zero trading activity is shown for each financial-instrument/Twitter-Filter combination. Based on these bin values, the Pearson's r values show the arithmetic mean of the correlations detected for the bins for each financial-instrument/Twitter-Filter combination.

			Mean Pearson's r correlation for:			
Filter ID	**Instrument**	**Filter**	**Positive sentiment**	**Negative sentiment**	**Net sentiment**	**Number of bins**
1	Apple, Inc. CFDs	$AAPL AND/OR “Apple”	-0.064	0.006	-0.021	235
2	Apple, Inc. CFDs	$AAPL	-0.129	-0.103	0.004	235
3	Amazon.com, Inc. CFDs	$AMZN AND/OR “Amazon”	-0.027	0.011	-0.022	189
6	Bank of America, Corp. CFDs	$BAC AND/OR “Bank of America”	0.082	0.032	0.004	231
8	Cisco Systems, Inc. CFDs	$CSCO AND/OR “Cisco”	0.114	0.034	0.091	176
15	Google, Inc. CFDs	$GOOG AND/OR “Google”	0.040	0.004	0.025	177
17	The Home Depot, Inc. CFDs	$HD AND/OR “Home Depot”	0.007	0.003	-0.063	143
20	IBM, Corp. CFDs	$IBM AND/OR “IBM”	0.005	0.003	0.058	189
22	Intel, Corp. CFDs	$INTC AND/OR “Intel”	0.105	-0.042	0.069	172
25	J.P. Morgan, Inc. CFDs	$JPM AND/OR “JPMorgan” AND/OR “JP Morgan”	0.148	0.089	-0.147	169
27	Coca-Cola, Co. CFDs	$KO AND/OR “Coca-Cola” AND/OR “Coca Cola”	0.052	-0.016	0.075	173
29	McDonald’s, Corp. CFDs	$MCD AND/OR “McDonald’s” AND/OR “McDonalds”	0.082	0.025	0.071	183
32	Microsoft, Corp. CFDs	$MSFT AND/OR “Microsoft”	-0.065	0.032	0.086	186
34	Oracle, Corp. CFDs	$ORCL AND/OR “Oracle”	0.030	-0.030	-0.056	165
36	FTSE100 Index CFDs	String-unfiltered UK Tweets	0.102	0.020	-0.031	147
37	FTSE100 Index Futures	String-unfiltered UK Tweets	Note A	Note A	Note A	Note A
38	S&P500 Index CFDs	String-unfiltered US Tweets	0.088	0.026	0.142	148
39	S&P500 Index Futures	String-unfiltered US Tweets	Note A	Note A	Note A	Note A
42	Wal-Mart, Inc. CFDs	$WMT AND/OR “Wal-Mart” AND/OR “Wal Mart”	-0.051	0.046	0.139	137

TABLE 7: RESULTS OF LINEAR REGRESSION ANALYSIS WITH BINNING BY NON-ZERO TRADING VOLUME

Table 7 shows the arithmetic mean Pearson’s r correlations between Twitter sentiment data and assets’ returns. **Note A**: There are no results for financial-instrument/Twitter-

Filter combinations whose assets were Futures – Futures markets do not shut, and periods of zero trading activity were not detected for the parameters of this experiment.

For the binning by non-zero trading volume configuration and for this experiment's parameters, the detected correlations show no or negligible relationships according to accepted interpretations of the values of Pearson's r[108,109].

6.2.3 Experiment 3: Binning by non-zero returns

The results of the experiments using binning by non-zero returns activity, as described in Chapter 5.4 are presented below in Table 8. Here, correlations are shown between hourly changes in the positive, negative and net sentiments of Tweets using hourly discretisation and 24-hour backward looking SMA smoothing and hourly returns of the financial assets.

The number of bins identifying chronological instances of non-zero returns is shown for each financial-instrument/Twitter-Filter combination. Based on these bin values, the Pearson's r values show the arithmetic mean of the correlations detected for the bins for each financial-instrument/Twitter-Filter combination.

			Mean Pearson's r correlation for:			
Filter ID	**Instrument**	**Filter**	**Positive sentiment**	**Negative sentiment**	**Net sentiment**	**Number of bins**
1	Apple, Inc. CFDs	$AAPL AND/OR “Apple”	-0.050	-0.015	-0.017	128
2	Apple, Inc. CFDs	$AAPL	-0.071	0.021	-0.03	128
3	Amazon.com, Inc. CFDs	$AMZN AND/OR “Amazon”	0.006	-0.004	0.005	142
6	Bank of America, Corp. CFDs	$BAC AND/OR “Bank of America”	0.152	0.076	-0.081	147
8	Cisco Systems, Inc. CFDs	$CSCO AND/OR “Cisco”	0.047	0.033	0.037	131
15	Google, Inc. CFDs	$GOOG AND/OR “Google”	0.028	0.014	-0.038	132
17	The Home Depot, Inc. CFDs	$HD AND/OR “Home Depot”	0.057	-0.017	-0.005	150
20	IBM, Corp. CFDs	$IBM AND/OR “IBM”	-0.012	0.007	-0.021	143
22	Intel, Corp. CFDs	$INTC AND/OR “Intel”	0.103	-0.093	0.115	129
25	J.P. Morgan, Inc. CFDs	$JPM AND/OR “JPMorgan” AND/OR “JP Morgan”	0.097	-0.058	-0.092	146
27	Coca-Cola, Co. CFDs	$KO AND/OR “Coca-Cola” AND/OR “Coca Cola”	0.029	0.014	-0.006	164
29	McDonald’s, Corp. CFDs	$MCD AND/OR “McDonald’s” AND/OR “McDonalds”	0.161	0.113	0.149	138
32	Microsoft, Corp. CFDs	$MSFT AND/OR “Microsoft”	-0.034	0.014	0.068	145
34	Oracle, Corp. CFDs	$ORCL AND/OR “Oracle”	0.035	-0.031	-0.047	139
36	FTSE100 Index CFDs	String-unfiltered UK Tweets	0.134	0.040	-0.003	167
37	FTSE100 Index Futures	String-unfiltered UK Tweets	0.042	0.008	0.016	426
38	S&P500 Index CFDs	String-unfiltered US Tweets	0.062	-0.006	0.059	131
39	S&P500 Index Futures	String-unfiltered US Tweets	0.041	0.029	0.061	490
42	Wal-Mart, Inc. CFDs	$WMT AND/OR “Wal-Mart” AND/OR “Wal Mart”	0.047	-0.028	0.128	122

Table 8: Results of linear regression analysis with binning by non-zero returns

Table 8 shows the arithmetic mean Pearson’s r correlations between Twitter sentiment data and assets’ returns. For the binning by non-zero returns configuration and for this experiment’s parameters, the detected correlations show no or negligible relationships according to accepted interpretations of the values of Pearson’s r[108,109].

6.2.4 Linear regression analysis summary

The results of the three linear regression analysis experiments, as detailed in Chapters 6.2.1, 6.2.2, and 6.2.3 show that there are no (or negligible) linear relationships between the social media data and the financial data. Granted, whilst the absence of significant relationships may be the by-product of the parameters used for these experiments, the absence of any strong correlations at the simple case of no time-shift indicate either that linear regression analysis is unable to identify indicative relationships between social media data and financial data in this study's dataset, or that the underlying relationships are nonlinear. It is indeed widely recognised that financial time-series are strongly nonlinear[88,89], as are the relationships between financial and social media data[9,110]. It is therefore not fruitful or necessary to explore the application of linear regression analysis to the problem at hand: the underlying datasets are either nonlinear, or this measure of dependency is not able to capture indicative relationships, or both.

Therefore, rather than exploring this measure of dependency further for example by fitting the linear regression analysis experiments' parameters until or if strong correlations are detected (which can be considered an exercise in parameter fitting just to produce a desired result), what is instead needed is the implementation an alternative measure of dependency without parameter fitting. Therefore, this study's dataset was next evaluated with a non-parameter optimised implementation of information theory (as discussed in Chapter 5.5) – a measure of dependency which can capture linear and nonlinear dependencies without model specification[85].

6.3 Information theory analysis results

The information theory analysis methodology, as described in Chapters 5.3.2 and 5.5 was used as a measure of dependency to explore the extent to which social media data leads the financial data without the limitations of assuming that any relationships are linear.

Chapter 5.5.2 describes criteria necessary to determine if social media data leads, rather than trails financial data. By using the notions of information surplus and time-shift as defined in Chapter 5.5.1, it is possible to quantify the extent to which social media data is more leading than trailing in relation to the financial data. Conceptually, this mechanism identifies when social media data carries more information about financial data ahead of time than at no leading time-shift to show which time-shifts, if any, result in social media data preceding financial data in a manner such that it is more leading than trailing.

If such time-shifts are detected for a particular financial-instrument/Twitter-Filter combination, they are tested for statistical-significance (as described in detail in Chapter 5.5.3) to assess which of these time-shifts are statistically-significant relative to randomly permeated data.

By applying this filtering mechanism to the nineteen financial-instrument/Twitter-Filter combinations which attract sufficient volumes of relevant messages, as listed in Table 5, the information theory analysis yields three possible outcomes denoting a null result:

1. Social media data is more trailing than leading. This is a null result, since the Twitter data is reactive rather than proactive;

2. Information surplus figures are negative. This is a null result, since Twitter data does not contain useful information relative to no time-shift;

3. Social media data is more leading than trailing, and information surplus figures are positive but the results are statistically-insignificant relative to randomly permeated data. This is a null result, since whilst the Twitter data contains useful information relative to no time-shift, and is proactive, it is not statistically-significant.

The only configuration under which the results are considered positive is:

- Social media data is more leading than trailing, and is therefore proactive and not reactive;

- Information surplus figures are positive, meaning Twitter data contains useful information relative to no time-shift;

- The results are statistically-significant to the 99% confidence level.

As with the results using linear regression analysis (see Chapter 6.2), the information theory analysis experiments were performed using hourly discretisation[a] of the data with 24-hour backwards-looking SMA smoothing (see Chapter 5.6.2.1).

[a] See Chapter 5.3 for an explanation of why hourly discretisation windows were used.

6.3.1 Null results for social media sentiment leading financial data

The financial-instrument/Twitter-Filter combinations which showed only null results for social media sentiment leading the financial data using the study's information theory analysis measure of dependency are summarised in the following table:

Filter ID	Instrument	Filter type
6	Bank of America, Corp. CFDs	Ticker-ID AND/OR Company Name
20	IBM, Corp. CFDs	Ticker-ID AND/OR Company Name
32	Microsoft, Corp. CFDs	Ticker-ID AND/OR Company Name
36	FTSE100 Index CFDs	UK Geographical
37	FTSE100 Index Futures	UK Geographical
38	S&P500 Index CFDs	US Geographical
42	Wal-Mart, Inc. CFDs	Ticker-ID AND/OR Company Name

TABLE 9: FINANCIAL-INSTRUMENT/TWITTER-FILTER COMBINATIONS FOR WHICH SOCIAL MEDIA SENTIMENT DOES NOT LEAD FINANCIAL DATA USING THE STUDY'S INFORMATION THEORY MEASURE OF DEPENDENCY

For each financial-instrument/Twitter-Filter combination listed in Table 9, hourly changes in the three sentiment types were considered independently to ascertain if changes in the positive sentiments, the negative sentiments and/or the net sentiments were able to lead the assets' hourly returns in a statistically-significant manner. In each case, it was also identified whether hourly changes in the Tweet message volumes were able to lead the assets' hourly returns or the absolute hourly returns in a statistically significant manner. As discussed in Chapter 5.6.2.1, the data were smoothed using a 24-hour backwards-looking SMA.

Details of the null results for each of the financial-instrument/Twitter-Filter combinations listed in Table 9 are given in the following subchapters. For each time-shift from 1-hour to 24-hours, the tables succinctly within the subchapters identify when:

- Social media data is more trailing than leading, denoted by "T>L";
- Information surplus figures are negative, denoted by "ISn";
- In cases where social media data is more leading than trailing, and information surplus figures are positive, the level of statistical-significance relative to randomly permeated data is given, denoted by "SS: xy.z%" (to demonstrate why it is a null result despite the aforementioned favourable conditions being identified).

Each table in Chapters 6.3.1.1 to 6.3.1.7 shows that for the time-shifts considered, hourly changes in Twitter sentiment or Twitter message volumes were not able to lead the assets' hourly returns. Tweets on the assets resulted in: instances where the social media data was more trailing than leading (T>L); instances where information surplus figures were negative (ISn); or instances where the social media data was more leading than trailing and information surplus figures were positive but the results were statistically-insignificant at the 99% confidence level, relative to randomly permeated data (SS: xy.z%).

The exception is for Bank of America, Corp. CFDs, for which Twitter message volumes were able to lead the asset's returns (as denoted by "Positive Result*" in Table 10) – this is discussed in further detail in Chapter 6.3.2, which reports the study's positive results.

6.3.1.1 Bank of America, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Bank of America, Corp. is a financial services provider and international bank headquartered in North Carolina, USA and is listed on the NASDAQ stock exchange with a market capitalisation of $179bn as at December 2014[a], and is the world's 318th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of commercial banking services such as Mergers & Acquisitions, Initial Public Offerings, market-making, and commercial debt finance.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$BAC" AND/OR "Bank of America", to capture Tweets mentioning Bank of America's Ticker-ID AND/OR the name of the company. In this manner, 208 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Bank of America, Corp. CFDs.

Time-Shift (hours) ↓. Parameter →	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	ISn	ISn	ISn	Positive Result*	Positive Result*
2	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
3	ISn T>L	ISn T>L	ISn T>L	Positive Result*	Positive Result*
4	ISn T>L	ISn T>L	ISn	Positive Result*	Positive Result*
5	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
6	ISn	ISn T>L	ISn	SS: 92.7%	SS: 94.7%
7	ISn	ISn T>L	ISn	SS: 97.5%	SS: 98.2%
8	ISn	ISn T>L	ISn	SS: 91.3%	SS: 93.4%
9	ISn	ISn T>L	ISn	SS: 87.8%	SS: 89.1%
10	SS: 55.3%	ISn T>L	ISn T>L	Positive Result*	SS: 98.6%
11	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
12	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
13	SS: 53%	ISn T>L	ISn	Positive Result*	Positive Result*
14	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
15	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
16	ISn	ISn T>L	ISn	SS: 97.5%	SS: 98%
17	ISn	ISn T>L	ISn	Positive Result*	Positive Result*
18	ISn	ISn T>L	ISn	SS: 87.2%	SS: 88.0%
19	ISn	ISn T>L	ISn	T>L	SS: 66.4%
20	ISn	ISn T>L	ISn T>L	SS: 87.8%	SS: 89.6%
21	ISn T>L	ISn T>L	ISn T>L	SS: 86.0%	SS: 87.6%
22	ISn T>L	ISn T>L	ISn T>L	ISn T>L	T>L
23	ISn T>L	ISn T>L	ISn T>L	ISn T>L	T>L
24	ISn T>L	ISn T>L	ISn T>L	T>L	T>L

TABLE 10: NULL RESULTS FOR BANK OF AMERICA VIA TICKER-ID AND/OR COMPANY NAME FILTERING

[a] http://uk.finance.yahoo.com/q/pr?s=BAC

6.3.1.2 IBM, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

IBM, Corp. a provider of information technology products and services worldwide, headquartered in New York, USA and is listed on the NASDAQ stock exchange with a market capitalisation of $150bn as at December 2014[a], and is the world's 10th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of IT infrastructure, business process services, cloud and technology services.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$IBM" AND/OR "IBM", to capture Tweets mentioning IBM's Ticker-ID AND/OR the name of the company. In this manner, 763 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of IBM, Corp. CFDs.

Time-Shift (hours) ↓. Parameter →	Positive sentiment vs. returns	Negative sentiment vs. returns	Net sentiment vs. returns	Tweet volume vs. returns	Tweet volume vs. absolute returns
1	T>L	ISn	ISn	ISn	ISn
2	T>L	ISn	ISn	SS: 58.0%	SS: 55.8%
3	ISn T>L	ISn	ISn	SS: 77.7%	SS: 77.4%
4	T>L	ISn	ISn T>L	SS: 65.5%	SS: 67.8%
5	ISn T>L	ISn	ISn	ISn	ISn
6	ISn T>L	ISn	ISn	SS: 96.2%	SS: 96.0%
7	ISn T>L	ISn	ISn T>L	ISn	ISn
8	ISn T>L	ISn	ISn	ISn	ISn
9	ISn T>L	ISn	ISn	SS: 92.8%	SS: 92.5%
10	ISn T>L	ISn	ISn	SS: 71.2%	SS: 68.4%
11	ISn T>L	ISn	ISn	SS: 72.0%	SS: 69.5%
12	ISn T>L	ISn	ISn	SS: 80.7%	SS: 79.4%
13	ISn T>L	ISn	SS: 54.6%	ISn	ISn
14	T>L	ISn	SS: 51.8%	SS: 86.2%	SS: 85.1%
15	ISn T>L	ISn	ISn	SS: 54.5%	SS: 52.0%
16	ISn T>L	ISn	ISn	ISn	ISn
17	ISn T>L	ISn	ISn	SS: 65.0%	SS: 61.1%
18	ISn T>L	ISn	ISn T>L	SS: 59.8%	SS: 56.9%
19	ISn T>L	ISn	ISn	ISn	ISn
20	T>L	ISn	ISn	ISn T>L	ISn
21	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
22	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
23	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
24	ISn T>L	SS: 64.3%	ISn T>L	ISn T>L	ISn T>L

TABLE 11: NULL RESULTS FOR IBM CFDS VIA TICKER-ID AND/OR COMPANY NAME FILTERING

[a] https://uk.finance.yahoo.com/q?s=IBM

6.3.1.3 Microsoft, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Microsoft, Corp. is a provider of software, IT services and IT hardware, headquartered in Washington, USA and is listed on the NASDAQ stock exchange with a market capitalisation of $376bn as at December 2014[a], and is the world's 4th highest-ranking company by brand popularity[111] at the time of writing. It is involved in the provision of IT products to consumer and business-to-business markets.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$MSFT" AND/OR "Microsoft", to capture Tweets mentioning Microsoft's Ticker-ID AND/OR the name of the company. In this manner, 3.9 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Microsoft, Corp. CFDs.

Time-Shift (hours) ↓. Parameter →	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	ISn T>L	SS: 63.4%	ISn	SS: 86.8%	SS: 86.5%
2	ISn T>L	SS: 64.8%	ISn T>L	SS: 86.8%	SS: 86.5%
3	ISn T>L	SS: 75.4%	ISn T>L	SS: 84.5%	SS: 82.3%
4	ISn T>L	SS: 91.5%	ISn	SS: 84.5%	SS: 82.3%
5	ISn T>L	SS: 68.9%	ISn	SS: 77.8%	SS: 76.0%
6	ISn T>L	SS: 82.5%	ISn	SS: 65.4%	SS: 64.9%
7	ISn T>L	SS: 97.6%	SS: 54.8%	SS: 95.7%	SS: 95.0%
8	ISn T>L	SS: 91.4%	ISn	SS: 85.1%	SS: 84.6%
9	ISn T>L	SS: 93.0%	ISn	SS: 60.5%	SS: 60.5%
10	ISn	SS: 97.3%	SS: 74.5%	SS: 97.2%	SS: 96.9%
11	ISn	SS: 60.0%	ISn	SS: 74.0%	SS: 71.8%
12	ISn	SS: 89.8%	ISn	SS: 53.4%	SS: 53.8%
13	ISn	SS: 79.0%	SS: 71.9%	SS: 85.1%	SS: 84.6%
14	ISn	SS: 69.9%	ISn	SS: 57.8%	SS: 58.3%
15	ISn	SS: 76.7%	ISn	SS: 50.3%	SS: 50.4%
16	ISn	SS: 62.1%	SS: 87.4%	ISn	ISn
17	ISn	ISn T>L	SS: 56.9%	ISn T>L	ISn
18	ISn	ISn	SS: 56.3%	ISn	ISn
19	ISn	ISn	SS: 66.9%	ISn	ISn
20	ISn T>L	ISn T>L	ISn	ISn T>L	ISn T>L
21	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
22	ISn T>L	ISn T>L	ISn	ISn T>L	ISn T>L
23	ISn T>L	ISn T>L	ISn T>L	ISn T>L	ISn T>L
24	ISn T>L	ISn T>L	ISn T>L	ISn T>L	ISn T>L

TABLE 12: NULL RESULTS FOR MICROSOFT CFDS VIA TICKER-ID AND/OR COMPANY NAME FILTERING

[a] https://uk.finance.yahoo.com/q?s=MSFT

6.3.1.4 FTSE100 Index CFDs, with social media source from string-unfiltered Tweets of UK origin

The FTSE100 Index is a share index of the largest 100 stocks in the UK.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any string-filtering using a geographical filter to capture Tweets tagged as originating from within the latitude and longitude coordinates encompassing the extremes of the United Kingdom. Note that this filtering method only detects those Tweets which contain a coordinate tag – a large proportion of Tweets are not tagged in this manner, however it is the most accurate method for filtering Tweets based on geographic origin.

In this manner, 4.7 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of FTSE100 Index CFDs.

Time-Shift (hours) ↓. Parameter →	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	ISn	SS: 62.5%	SS: 92.3%	ISn	ISn T>L
2	ISn	SS: 91.0%	SS: 98.1%	ISn	ISn T>L
3	ISn	SS: 77.8%	SS: 89.4%	ISn	ISn
4	SS: 51.9%	SS: 81.8%	SS: 56.4%	ISn T>L	ISn T>L
5	ISn	SS: 98.9%	SS: 76.6%	ISn	ISn T>L
6	ISn	SS: 55.1%	ISn	ISn T>L	ISn T>L
7	ISn	SS: 90.3%	ISn T>L	ISn T>L	ISn T>L
8	SS: 54.9%	SS: 71.0%	ISn	ISn T>L	ISn T>L
9	ISn	ISn	SS: 58.4%	ISn T>L	ISn T>L
10	ISn	SS: 68.3%	ISn T>L	ISn T>L	ISn T>L
11	ISn	ISn	ISn T>L	ISn T>L	ISn T>L
12	ISn	ISn	ISn	ISn T>L	ISn T>L
13	ISn	ISn	ISn	ISn T>L	ISn T>L
14	ISn	ISn	ISn T>L	ISn T>L	ISn T>L
15	ISn	ISn	ISn	ISn T>L	ISn T>L
16	ISn T>L	ISn	SS: 51.6%	ISn T>L	ISn T>L
17	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
18	ISn	ISn	ISn	ISn T>L	ISn T>L
19	ISn T>L	ISn	SS: 54.6%	ISn T>L	ISn T>L
20	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
21	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
22	ISn T>L	ISn	SS: 57.1%	ISn T>L	ISn T>L
23	ISn T>L	ISn	SS: 71.2%	ISn T>L	ISn T>L
24	ISn T>L	ISn	ISn	ISn T>L	ISn T>L

Table 13: Null results for FTSE100 CFDs via UK-geographical filtering

6.3.1.5 FTSE100 Index Futures, with social media source from string-unfiltered Tweets of UK origin

The same collection process was used for this financial-instrument/Twitter-Filter combination as the FTSE100 Index Futures (Chapter 6.3.1.4). Thus, 4.7 million Tweets were analysed to ascertain the extent to which they can lead the hourly returns of FTSE100 Index Futures.

Time-Shift (hours) ↓. **Parameter →**	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	ISn	ISn	SS: 92.3%	ISn	ISn
2	ISn	ISn	SS: 89.0%	ISn	ISn
3	ISn T>L	ISn	SS: 65.6%	ISn	ISn
4	ISn T>L	ISn	SS: 67.6%	ISn	ISn T>L
5	ISn	ISn	SS: 83.2%	ISn T>L	ISn T>L
6	ISn	SS: 57.4%	SS: 81.4%	ISn	ISn
7	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
8	ISn	ISn	SS: 73.2%	ISn T>L	ISn T>L
9	ISn	ISn	ISn	ISn T>L	ISn T>L
10	ISn	SS: 72.4%	ISn	ISn T>L	ISn T>L
11	ISn	ISn	ISn	ISn T>L	ISn T>L
12	ISn	ISn	ISn T>L	ISn T>L	ISn T>L
13	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
14	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
15	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
16	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
17	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
18	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
19	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
20	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
21	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
22	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
23	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L
24	ISn T>L	ISn	ISn T>L	ISn T>L	ISn T>L

TABLE 14: NULL RESULTS FOR FTSE100 FUTURES VIA UK-GEOGRAPHICAL FILTERING

6.3.1.6 S&P500 Index CFDs, with social media source from string-unfiltered Tweets of US origin

Standard and Poor’s 500 Index is a capitalisation-weighted index of the largest 500 stocks in the USA.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter’s 10% Gardenhose Feed without any string-filtering using a geographical filter to capture Tweets tagged as originating from within the latitude and longitude coordinates encompassing the extremes of contiguous United States of America. Note that this filtering method only detects those Tweets which contain a coordinate tag – a large proportion of Tweets are not tagged in this manner, however it is the most accurate method for filtering Tweets based on geographic origin.

In this manner, 18.7 million Tweets were filtered in during this study’s 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of S&P500 Index CFDs.

Time-Shift (hours) ↓. Parameter →	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	ISn T>L	ISn	ISn	ISn	ISn
2	T>L	ISn	ISn	ISn	ISn
3	T>L	ISn	ISn	ISn	ISn
4	T>L	ISn	ISn	ISn	ISn
5	SS: 88.5%	ISn	ISn T>L	ISn	ISn
6	SS: 88.1%	ISn	ISn	ISn	ISn
7	SS: 88.5%	ISn	ISn T>L	ISn T>L	ISn
8	SS: 81.4%	ISn	ISn T>L	ISn	ISn
9	SS: 69.9%	ISn	ISn T>L	ISn	ISn
10	T>L	ISn	ISn T>L	ISn	ISn
11	SS: 83.4%	SS: 78.8%	ISn T>L	ISn	ISn
12	SS: 80.1%	SS: 74.3%	ISn	ISn	ISn
13	ISn T>L	SS: 84.8%	ISn	ISn	ISn
14	ISn T>L	SS: 85.1%	ISn	ISn T>L	ISn T>L
15	ISn T>L	SS: 87.0%	ISn	ISn T>L	ISn T>L
16	ISn T>L	SS: 80.4%	SS: 77.4%	ISn T>L	ISn T>L
17	ISn T>L	SS: 82.6%	SS: 60.6%	ISn T>L	ISn T>L
18	ISn T>L	SS: 71.9%	ISn	ISn T>L	ISn T>L
19	ISn T>L	SS: 82.6%	SS: 75.3%	ISn T>L	ISn T>L
20	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
21	ISn T>L	ISn	ISn	ISn T>L	ISn T>L
22	ISn T>L	ISn	SS: 56.4%	ISn T>L	ISn T>L
23	ISn T>L	SS: 70.6%	SS: 55.3%	ISn T>L	ISn T>L
24	ISn T>L	SS: 63.4%	ISn	ISn T>L	ISn T>L

Table 15: Null results for S&P500 CFDs via US-geographical filtering

6.3.1.7 Wal-Mart, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Wal-Mart, Inc. is an operator of retail stores in the US and internationally, headquartered in Arkansas, USA and is listed on the New York Stock Exchange with a market capitalisation of $270bn as at December 2014[a], and is the world's 9th highest-ranking company by brand popularity[111] at the time of writing. It is involved in the provision of groceries, home products and financial services to retail customers.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$WMT" AND/OR "Wal-Mart" AND/OR "Wal Mart", to capture Tweets mentioning Wal-Mart's Ticker-ID AND/OR the name of the company. In this manner, 720 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Wal-Mart, Inc. CFDs.

Time-Shift (hours) ↓. Parameter →	**Positive sentiment vs. returns**	**Negative sentiment vs. returns**	**Net sentiment vs. returns**	**Tweet volume vs. returns**	**Tweet volume vs. absolute returns**
1	SS: 54.1%	T>L	SS: 78.6%	SS: 63.7%	SS: 57.8%
2	ISn	ISn T>L	T>L	SS: 88.8%	SS: 85.5%
3	SS: 50.2%	T>L	ISn T>L	SS: 92.5%	SS: 89.5%
4	SS: 58.8%	SS: 74.4%	SS: 68.8%	SS: 75.5%	SS: 69.2%
5	ISn	SS: 78.8%	ISn T>L	ISn	ISn
6	ISn	T>L	ISn T>L	SS: 87.8%	SS: 84.3%
7	ISn	SS: 98.8%	ISn T>L	ISn	ISn
8	ISn	SS: 97.3%	ISn T>L	ISn T>L	ISn T>L
9	ISn	SS: 85.6%	ISn T>L	SS: 88.8%	SS: 85.1%
10	SS: 53.7%	SS: 97.3%	ISn T>L	SS: 81.2%	SS: 78.7%
11	ISn	SS: 91.4%	ISn T>L	SS: 49.2%	ISn
12	ISn	SS: 80.5%	ISn T>L	SS: 88.2%	SS: 84.9%
13	ISn	SS: 94.2%	ISn T>L	SS: 92.5%	SS: 91.4%
14	ISn	SS: 97.7%	ISn T>L	SS: 88.8%	SS: 85.1%
15	ISn	SS: 88.5%	ISn T>L	SS: 82.8%	SS: 77.6%
16	ISn	SS: 71.2%	ISn T>L	SS: 83.3%	SS: 78.2%
17	ISn	SS: 80.3%	ISn T>L	SS: 84.3%	SS: 79.5%
18	ISn	ISn T>L	ISn T>L	SS: 77.7%	SS: 71.3%
19	ISn	SS: 71.0%	ISn T>L	SS: 71.8%	SS: 64.7%
20	ISn	SS: 79.0%	ISn T>L	SS: 93.8%	SS: 92.6%
21	ISn	T>L	ISn T>L	SS: 97.0%	SS: 95.7%
22	ISn	ISn T>L	ISn T>L	SS: 82.3%	SS: 76.9%
23	ISn	ISn T>L	ISn T>L	SS: 70.3%	SS: 62.7%
24	SS: 58.8%	ISn T>L	ISn T>L	SS: 91.2%	SS: 91.4%

TABLE 16: NULL RESULTS FOR WAL-MART CFDS VIA TICKER-ID AND/OR COMPANY NAME FILTERING

[a] https://uk.finance.yahoo.com/q?s=WMT

6.3.1.8 Summary of null results (using information theory analysis)

In summary, the information theory analysis methodology as described in Chapters 5.3.2 and 5.5, was used as a measure of dependency to explore the extent to which social media data leads the financial data. As with the experiments using linear regression analysis as a measure of dependency, this was performed using the 24-hour backwards-looking SMA smoothing condition as described in Chapter 5.6.2.1. For each time-shift from 1-hour to 24-hours, the methodology tested whether the social media is more leading than trailing; whether a positive information surplus is detected; and whether the results are statistically-significant.

This methodology identified six financial-instrument/Twitter-Filter combinations for which neither Twitter sentiment nor Twitter message volumes were able to lead the assets' hourly returns. However, one further financial-instrument/Twitter-Filter combination (Bank of America, Corp. CFDs) was identified for which Twitter sentiment showed null results, but Twitter message volumes were able to lead the asset's hourly returns – this is discussed in further detail in Chapter 6.3.2, which reports the study's positive results.

6.3.2 Positive results for social media sentiment leading financial data

The financial-instrument/Twitter-Filter combinations which showed positive results for social media sentiment leading the financial data using the study's information theory analysis measure of dependency are discussed in this chapter.

The configuration under which the results are considered positive is:

- Social media data is more leading than trailing, and is therefore proactive and not reactive;
- Information surplus figures are positive, meaning Twitter data contains useful information relative to no time-shift;
- The results are statistically-significant to the 99% confidence level.

As with the results using linear regression analysis (see Chapter 6.2), the information theory analysis experiments were performed using hourly discretisation[a] of the data with 24-hour backwards-looking SMA smoothing (see Chapter 5.6.2.1).

According to the methodology described in Chapter 5.5.4, the study's experiments were also repeated to consider $\Delta_{\text{message volume}}$ against Δ_{price} (the returns), as well as $\Delta_{\text{message volume}}$ against $|\Delta_{\text{price}}|$ (the absolute returns) as an echo of past studies which compare Google Search Trends[13-15] and Yahoo! search engine data[16] message volumes with financial market performance. These experiments allow for the identification of the extent to which hourly changes in Twitter message sentiments lead securities' hourly returns over and above what is attainable by the evaluation of hourly changes in Twitter message volumes.

A summary of the positive results is given in Chapter 6.3.2.1, and details for each financial-instrument/Twitter-Filter combination are given in the Appendix (Chapter 11.1).

[a] See Chapter 5.3 for an explanation of why hourly discretisation windows were used.

6.3.2.1 Summary of positive results for social media leading financial data

For each financial-instrument/Twitter-Filter combination, Table 17 lists the leading time-shifts which result in the largest statistically-significant information surplus values for social media sentiment leading the financial data. In each case, the results for experiments investigating the dependencies between message volumes and returns, as well as message volumes vs. absolute returns, are also presented.

Filter ID	**Instrument**	**Filter type**	**Mean message volume per minute**	**Largest statistically-significant information surplus from sentiment vs. returns**	**Largest statistically-significant information surplus from message volume vs. returns**	**Largest statistically-significant information surplus from message volume vs. absolute returns**
1	Apple, Inc. CFDs	Ticker-ID AND/OR Company Name	126.7	0.14%	N/A	N/A
2	Apple, Inc. CFDs	Ticker-ID	1.8	3.35%	0.89%	0.94%
3	Amazon.com, Inc. CFDs	Ticker-ID AND/OR Company Name	123.1	3.47%	N/A	N/A
6	Bank of America, Corp. CFDs	Ticker-ID AND/OR Company Name	1.6	N/A	0.60%	0.65%
8	Cisco Systems, Inc. CFDs	Ticker-ID AND/OR Company Name	4.0	2.77%	N/A	N/A
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184.0	2.64%	N/A	N/A
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	1.9	2.81%	2.02%	2.23%
22	Intel, Corp. CFDs	Ticker-ID AND/OR Company Name	12.9	1.41%	N/A	0.52%
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	3.94%	1.21%	1.37%
27	Coca-Cola, Co. CFDs	Ticker-ID AND/OR Company Name	24.8	0.72%	N/A	N/A
29	McDonald’s, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	1.90%	N/A	N/A
34	Oracle, Corp. CFDs	Ticker-ID AND/OR Company Name	5.0	0.36%	N/A	N/A
39	S&P500 Futures	US Geographical	**142.7***	2.46%	N/A	N/A

Filter ID	**Instrument**	**Leading time-shift corresponding to the largest information surplus from sentiment vs. returns**	**Sentiment type corresponding to the largest information surplus**	**Number of statistically-significant leading information surplus time-shifts from sentiment vs. returns**	**Number of statistically-significant leading information surplus time-shifts from message volume vs. returns**	**Number of statistically-significant leading information surplus time-shifts from message volume vs. absolute returns**
1	Apple, Inc. CFDs	10 hours	Negative	2	N/A	N/A
2	Apple, Inc. CFDs	14 hours	Negative	2	5	9
3	Amazon.com, Inc. CFDs	20 hours	Net	30	N/A	N/A
6	Bank of America, Corp. CFDs	N/A	**N/A**	**N/A**	8	8
8	Cisco Systems, Inc. CFDs	13 hours	Net	15	N/A	N/A
15	Google, Inc. CFDs	14 hours	Net	14	N/A	N/A
17	The Home Depot, Inc. CFDs	11 hours	Positive	8	4	4
22	Intel, Corp. CFDs	1 hour	Negative	2	N/A	2
25	J.P. Morgan, Inc. CFDs	12 hours	Positive	2	15	14
27	Coca-Cola, Co. CFDs	8 hours	Positive	13	N/A	N/A
29	McDonald’s, Corp. CFDs	13 hours	Net	7	N/A	N/A
34	Oracle, Corp. CFDs	1 hour	Net	1	N/A	N/A
39	S&P500 Futures	22 hours	Net	1	N/A	N/A

TABLE 17: SUMMARY OF THE POSITIVE RESULTS FOR SOCIAL MEDIA’S ABILITY TO LEAD SECURITIES’ RETURNS AHEAD OF TIME

For each financial-instrument/Twitter-Filter combination, Table 17 shows:

- The search characteristics of the Twitter filters, and corresponding mean message volumes over the study's entire dataset;
- The largest statistically-significant information surplus values for each parameter type (sentiment, message volume vs. returns and message volume vs. absolute returns). This describes the largest amount of information that Twitter data contains ahead of time relative to no time-shift (see Chapter 5.5.1 for a full explanation of information surplus). The time-shifts at which these statistically-significant information surplus values are detected are also given (see Chapter 5.3.2 for a full explanation of time-shifts);
- In the case of the sentiment parameter, the sentiment type (positive, negative or net) which provided the largest statistically-significant information surplus values is also given (see Chapter 5.1 for a full explanation of sentiment types);
- The total number of statistically-significant time-shifts for each parameter type for which the social media data were able to lead the financial data in a statistically-significant manner.

Table 17 shows twelve financial-instrument/Twitter-Filter combinations for which Tweet sentiments were able to lead the financial returns data. For some assets, message volumes also showed abilities to lead the financial returns data, but these abilities are weaker than with the sentiment experiments.

Note that as mentioned in Chapter 6.3.1.1, Tweet sentiments on Bank of America, Corp. were not able to lead the returns of the company's CFDs – however, Tweet message volumes were. This is discussed further in the Appendix (Chapter 11.1.4).

Details of the positive results for each of the financial-instrument/Twitter-Filter combinations listed in Table 17 are given in the Appendix (Chapter 11.1). However, summarisations of the positive results for each of the twelve financial-instrument/Twitter-Filter combinations for which Tweet sentiments were able to lead the financial returns data, are shown in Figure 19. This figure shows the leading time-shifts which resulted in the largest statistically-significant information surplus values out of the three sentiment types (positive, negative or net). This figure is split into two charts of six assets each for clarity of visualisation.

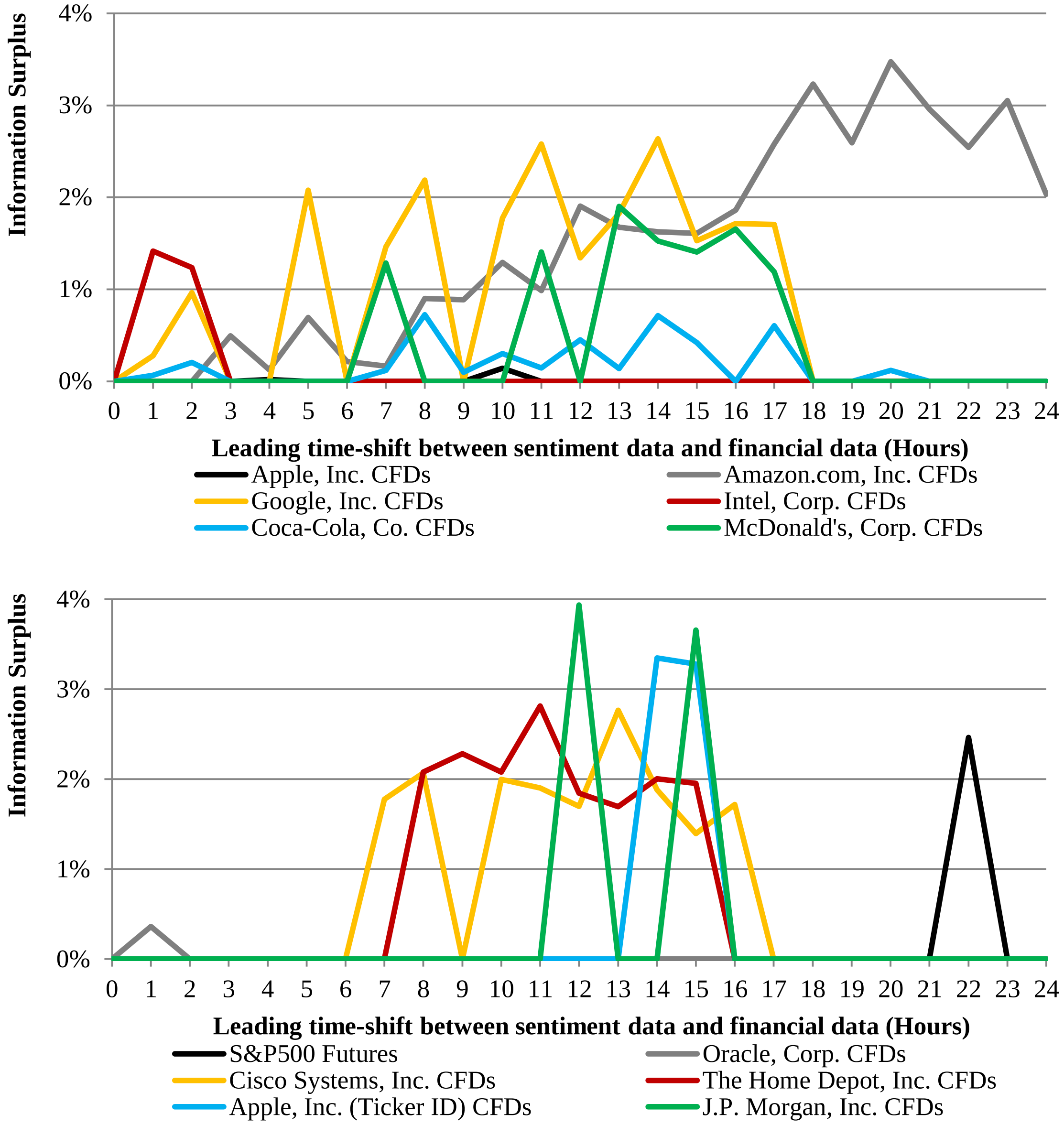


FIGURE 19: TIME-SHIFTS BETWEEN HOURLY CHANGES IN TWITTER MESSAGE SENTIMENTS AND SECURITIES' RETURNS WHICH RESULT IN THE LARGEST STATISTICALLY-SIGNIFICANT INFORMATION SURPLUS VALUES

6.3.2.2 Sentiment outperforms volumes of messages in leading assets' returns

Figure 19 shows the performance of Twitter sentiment leading financial returns. To compare this to the performance of Twitter message volumes leading financial returns, and absolute financial returns, consider Figure 20, below.

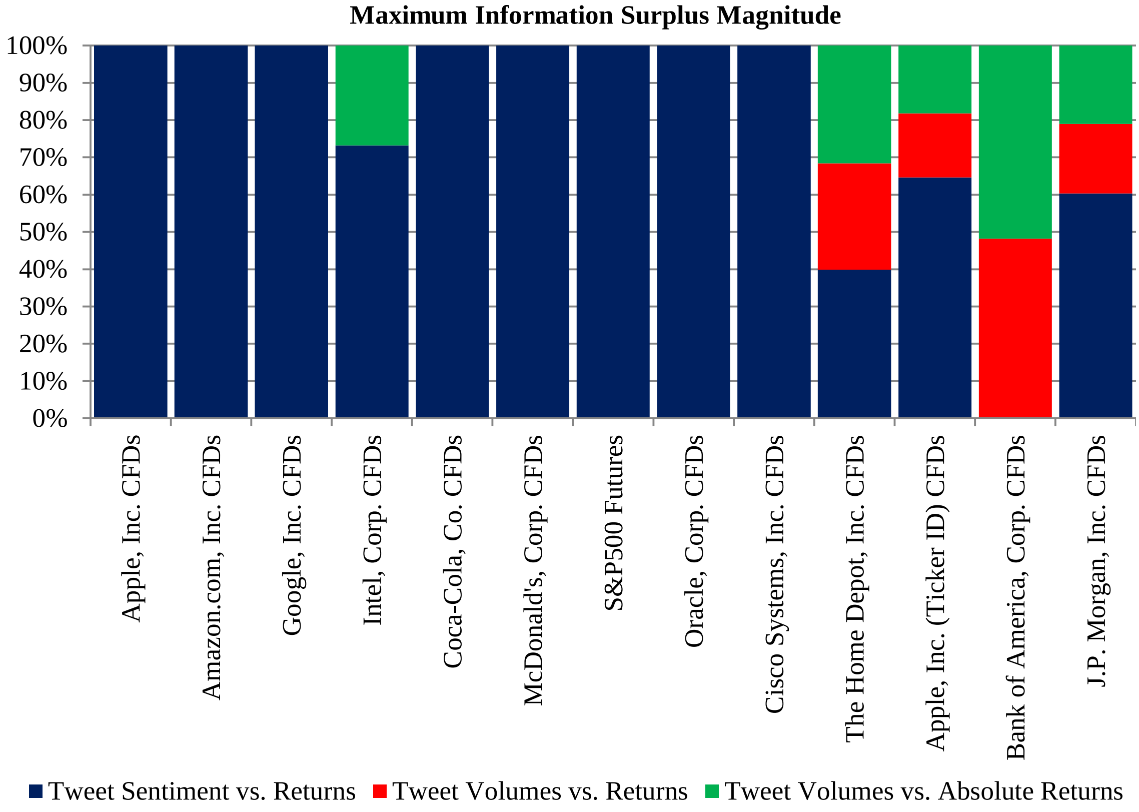


FIGURE 20: THE RATIO OF LARGEST INFORMATION SURPLUS VALUES FROM THE ANALYTICS OF SENTIMENT RELATIVE TO THE ANALYTICS OF TWEET MESSAGE VOLUMES

Figure 20 shows the ratio of the largest information surplus figures for experiments which evaluate Twitter sentiment vs. returns, relative to experiments which evaluate message volumes vs. returns, and message volumes vs. absolute returns. This figure shows that the experiments which measure message sentiment result in proportionally larger maximum information surplus values relative to the experiments which only measure message volume.

The study’s positive results therefore show that comparatively larger information surplus values are attainable from social media sentiment rather than from social media message volumes.

To further support this point, Figure 21 shows that the analytics of social media sentiment results in proportionally more statistically-significant information surplus instances than the number of statistically-significant information surplus instances observed from the analytics of social media message volumes.

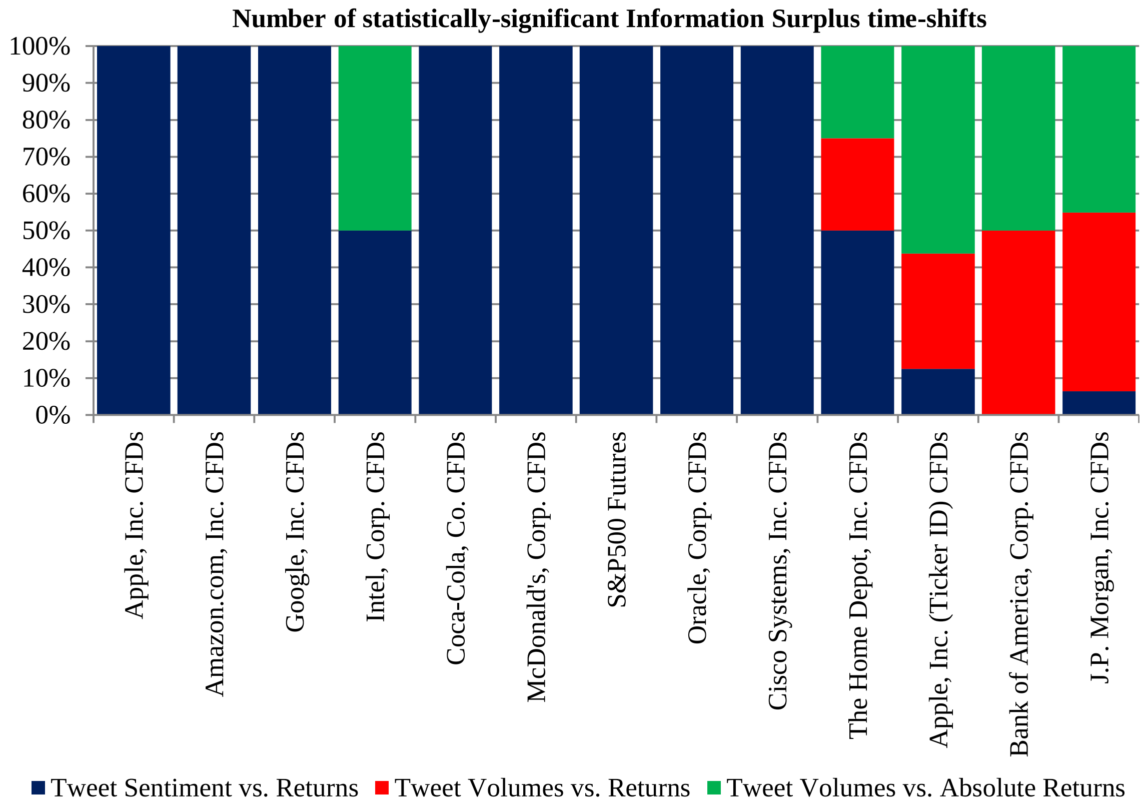


FIGURE 21: THE RATIO OF INFORMATION SURPLUS INSTANCES FROM THE ANALYTICS OF SENTIMENT RELATIVE TO THE ANALYTICS OF TWEET MESSAGE VOLUMES

Figure 21 shows that hourly changes in Tweet message sentiments (blue bars) led the asset’s hourly returns more often than hourly changes in Tweet message volumes, whether these volumes are evaluated against hourly returns (red bars) or absolute hourly returns (green bars).

As discussed earlier, there is however one exception: the sentiments of Tweets on Bank of America, Corp. did not show instances of leading the returns of the firm’s CFDs, whilst the message volumes of the Tweets did (see Chapter 11.1.4).

6.3.3 Generalisation of positive results for leading financial data with social media data using information theory analysis

Generalisations can be made from the positive results (Chapter 6.3.2) of the information theory analysis (method discussed in Chapter 5.5) under the study's constraints. From the analysis of 10% of Tweets from Twitter's network (as discussed in Chapter 2.1) as collected over a 3-month period (as justified in Chapter 5.2), the constraints are:

1. The analysis of Tweets as discretised into hourly windows (as justified in Chapter 5.3) using hourly time-shifts of up to 24-hours (as discussed in Chapter 5.3.2.2);

2. The analysis of the dependency between Twitter data and financial data using discretisation windows on the hour (as detailed in Chapter 5.3);

3. The use of Sturges' Method for histogram binning in the calculations of mutual information (as justified in Chapter 5.3.2.1);

4. The analysis of Tweet message sentiments and volumes as smoothed by a backwards-looking simple moving average window of 24-hours (as justified in Chapter 5.6.2.1);

LIST 1: THE STUDY'S CONSTRAINTS FOR ITS INFORMATION THEORY ANALYSIS EXPERIMENTS

Generalisations which can be inferred from the positive results of the information theory analysis are listed below:

1. The results of these experiments indicate that Tweet message sentiment adds information to Tweet message volumes for some assets from the pool of assets considered in this research under the study's constraints. The additional gains from sentiment, over message volumes, are detailed in Chapter 6.3.2.2.

2. The results of these experiments indicate that as the average message volume per financial-instrument/Twitter-Filter combinations relating to individual companies increases, so does the number of time-shifts for which sentiment leads the financial data for those companies. I.e., this study's experiments show that a greater message volume indicated the possibility that social media

sentiment is more predictive for individual companies. Consider Figure 22, below:

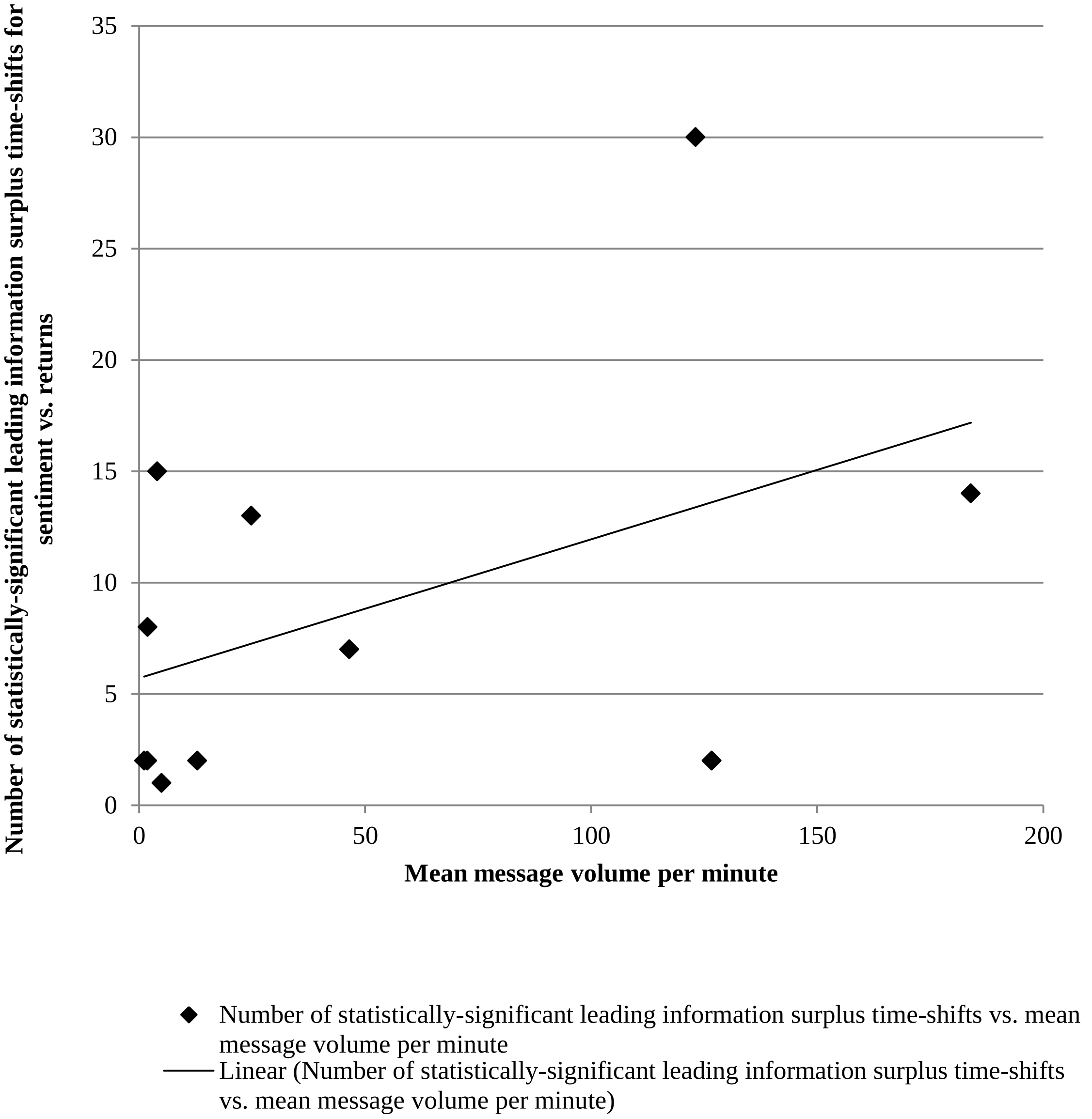


FIGURE 22: RELATIONSHIP BETWEEN MEAN MESSAGE VOLUME PER MINUTE AND THE NUMBER OF STATISTICALLY-SIGNIFICANT LEADING INFORMATION SURPLUS TIME-SHIFTS FOR TWITTER SENTIMENT VS. ASSET RETURNS

Figure 22 shows under the constraints of the study’s information theory experiments, a positive relationship exists between the mean message volume per minute per financial-instrument/Twitter-Filter combination, and the number of statistically-significant leading information surplus time-shifts for sentiment vs. returns. Note: Figure 22 only includes data for Twitter-Filters referring to specific publically-traded companies, and not the S&P500 data. These data give a Pearson's r correlation of 0.459, which denotes a moderate-to-strong positive relationship according to accepted interpretations of the values of Pearson's r[108,109].

3. The results of these experiments indicate that as the average message volume per financial-instrument/Twitter-Filter relating to individual companies increases, the larger the time-shift is for which the largest information surplus is detected. The study's experiments therefore show that a greater message volume indicated the possibility that social media sentiment is predictive more in advance (i.e., further ahead of time). Consider Figure 23, below:

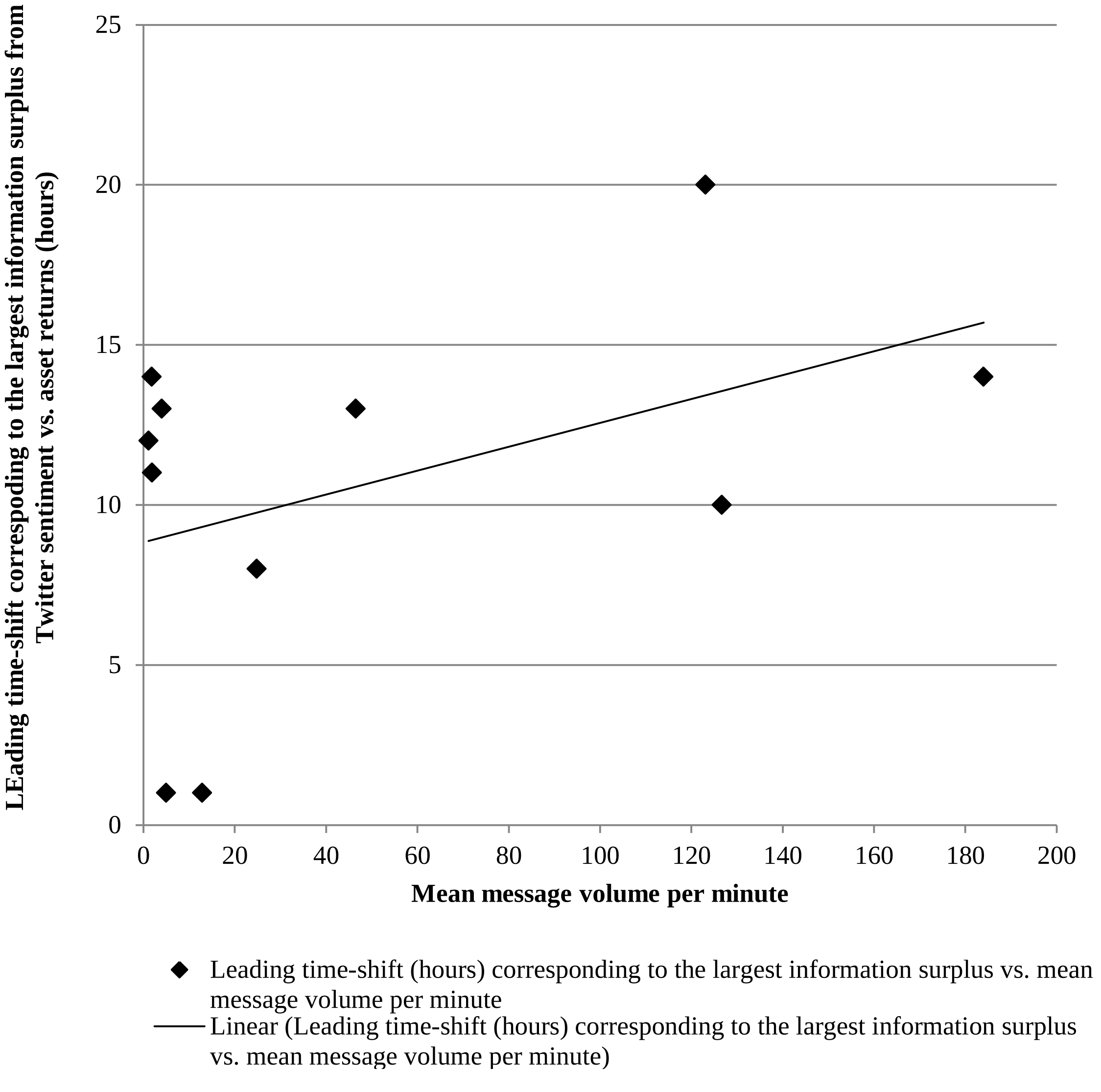


FIGURE 23: RELATIONSHIP BETWEEN MEAN MESSAGE VOLUME PER MINUTE AND THE LARGEST INFORMATION SURPLUS FROM TWITTER SENTIMENT VS. ASSET RETURNS

Figure 23 shows that under the constraints of the study's information theory experiments, a positive relationship exists between the mean message volume per minute per financial-instrument/Twitter-Filter combination, and the leading time-shift corresponding to the largest information surplus from Twitter sentiment vs. asset returns. Note: Figure 23 only includes data for Twitter-Filters referring to specific publically-traded companies, and not the S&P500 data. These data give a Pearson's r correlation of 0.431, which denotes a moderate-to-strong positive relationship according to accepted interpretations of the values of Pearson's r[108,109].

The aforementioned constraints (List 1) and generalisations would be conducive to permitting practical medium or high frequency trading from Twitter sentiment data, depending on one's definition of these terms. From a 2009 survey[112] of 202 traders from hedge-funds, investment advisory or financial consulting firms, 86% of respondents defined "high-frequency trading" as referring to holding periods of 24-hours or less. Considering this study's constraints applied to the sub 24-hour time-shift scope, it can be stated that the positive results of this research are applicable to practical applications of trading from Twitter sentiment within the predominating definition of high-frequency trading.

Furthermore, as is discussed in Chapter 6.1, with consideration for the best practical application of the study's findings, the research was centred on the analysis of Tweets on financial instruments which attract a sufficient message volume rate to fully utilise Twitter's API call capacity.

The robustness of the study's results against parameter variation is tested in Chapter 7.

7 ROBUSTNESS OF RESULTS

This chapter details the robustness of the study's results from the information theory analysis experiments against parameter variation.

Under this study's constraints (List 1), a series of generalisations of the results were made (as listed in Chapter 6.3.3), showing that:

1. Message sentiment adds information over what is attainable from message volumes;
2. A greater message volume indicated the possibility that social media sentiment is more predictive for individual companies;
3. A greater message volume indicated the possibility that social media sentiment is predictive more in advance (i.e., further ahead of time) for individual companies.

LIST 2: LIST OF GENERALISATIONS OF THE STUDY'S RESULTS UNDER ITS CONSTRAINTS

The above statements point to the question: do these generalisations change significantly with variations in the constraints of the study (as detailed in List 1)? Specifically, by instituting parameter variation:

1. For point 1, in List 2:
   - Is the conclusion strengthened for an asset where minimal information is added by sentiment to message volume?
   - Is the conclusion weakened for an asset where maximal information is added by sentiment to message volume?
   - Is there significant variation for an asset where the mean amount of information is added by sentiment to message volume across all assets from the study?

2. For point 2, in List 2:

    - Is the conclusion strengthened for an asset where minimal message volume is observed across the dataset?
    - Is the conclusion weakened for an asset where maximal message volume is observed across the dataset?
    - Is there significant variation for an asset where the mean message volume is observed from all assets across the dataset?

3. For point 3, in List 2:

    - Is the conclusion strengthened for an asset where minimal message volume is observed across the dataset?
    - Is the conclusion weakened for an asset where the maximal message volume is observed across the dataset?
    - Is there significant variation for an asset where the mean message volume is observed from all assets across the dataset?

LIST 3: EXPLORING THE STRENGTHENING OR WEAKENING OF THE STUDY'S GENERALISED RESULTS UNDER ITS CONSTRAINTS

This chapter therefore answers the question of the extent to which the variation in the study's constraints (as detailed in List 1) affects the study's generalised results (as detailed in List 2) according to the possible outcomes of this exercise (as detailed in List 3).

The following subchapters detail the effects of the variation of the study's constraints, where applicable, on an independent variable-by-variable basis. These effects are reported with reference to the study's generalised results as detailed in List 2.

7.1 Sensitivity to discretisation window size

As discussed in Chapters 5.3 and 5.6.1, the choice of discretisation frequency in the financial services industry is often ad-hoc, typically dictated by the observation intervals of the available data[79]. The development of SocialSTORM[57], which provided preliminary results for this study, used financial data which was not available to resolutions smaller than hourly[a][80]. Since these preliminary investigations using hourly discretisation windows showed support for the existence of dependencies between social media and financial data[80-82], the study's experiments were also performed using this discretisation window size.

Testing the sensitivity of the study's information theory analysis results to discretisation window size is useful in identifying whether varying the hourly discretisation size as selected based on the resolution of the data during the development of SocialSTORM[57] produces significant differences to the results.

Increasing the size of the discretisation window beyond 1-hour would have an effect of diluting the feature density of the study's dataset and its results. This would be due to the effect of averaging caused by the process of the allocation of data to larger windows by way of mean averaging during the data-discretisation process (as described in Chapter 5.3). Instead, what is of interest is if whether the allocation of the study's dataset to discretisation windows of a higher resolution than 1-hour would produce significant results variation. Given that the financial dataset used in the TCF (see Chapter 5.6.1) had a common highest resolution of 5-minutes, the sensitivity of the study's information theory analysis results are tested by repeating the experiments using this highest-resolution discretisation window size common to the dataset.

Thus, the analysis theory described in Chapter 5.3.2 and implemented via the methodology described in Chapter 5.5 was repeated on key Financial-instrument/Twitter-Filter combinations (as detailed in List 3) using the 5-minute (rather than 1-hour) discretisation window size. The effect of using a 5-minute discretisation window is shown diagrammatically below in Figure 24.

[a] Financial data used for the preliminary investigation was sourced from Thomson Reuters and from Fulcrum Asset Management, and was discretised to hourly windows due to the unavailability of higher-resolution data at the time.

The study's main results are based on the discretisation of continuous social media and financial data by way of arithmetic mean averaging (as discussed in Chapter 5.3). In this study, the social media and financial data were discretised into hourly windows (as denoted by the yellow arrows). To test the robustness of the study's results, its information theory analysis experiments were repeated on key Financial-instrument/Twitter-Filter combinations (as detailed in List 3) using discretisation windows of 5-minutes (as denoted by the blue arrows).

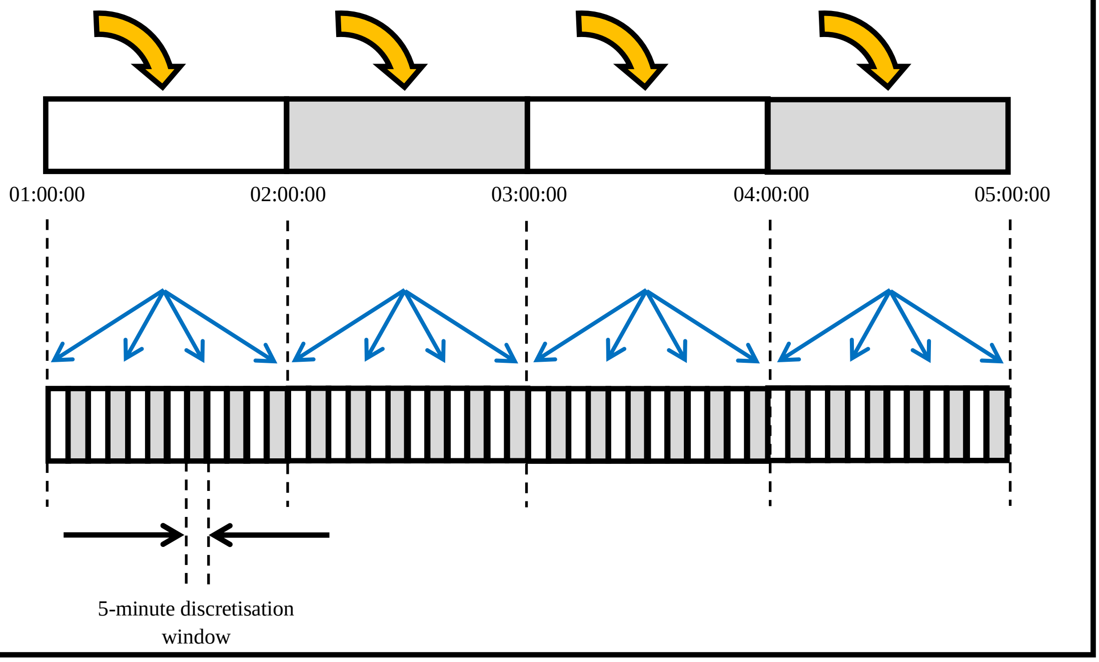


Figure 24: Example of the implementation of the 5-minute data discretisation window

The results of this robustness experiment are presented in the subsequent subchapters, and in such manner address the first parameter variation point in List 1.

7.1.1 Sensitivity to discretisation window size: robustness results for message sentiment adding information to what is available from message volumes

The Financial-instrument/Twitter-Filter combinations selected for this sensitivity experiment, according to the criteria in List 3, are detailed below.

Financial-instrument/Twitter-Filter combination A:

- For which a minimal quantity of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;

  - Information added by sentiment over message volume when evaluated against asset returns: 0.79%;
  - Information added by sentiment over message volume when evaluated against absolute asset returns: 0.58%.

Financial-instrument/Twitter-Filter combination B:

- For which maximal information is added by message sentiment over what is attainable from message volumes in the study’s findings as listed in Chapter 6.3.2:
  - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
  - Information added by sentiment over message volume when evaluated against asset returns: 2.73%;
  - Information added by sentiment over message volume when evaluated against absolute asset returns: 2.57%.

Financial-instrument/Twitter-Filter combination C:

- The aim is to identify an asset for which a mean amount of information is added by message sentiment over what is attainable from message volumes. However, as detailed in Table 17, only three assets were identified for which both social media volume and sentiments were able to lead financial returns. Given that two of the three assets are already being explored (points B, and C, above), it is not possible to identify a mean. Therefore, the remaining third asset is used for this experiment since it attracts an intermediate additional amount of information from message sentiment over what is attainable from message volumes.
- For which an intermediate amount of information is added by message sentiment over what is attainable from message volumes in the study’s findings as listed in Chapter 6.3.2:
  - Apple, Inc. CFDs from the Ticker-ID only Twitter Filter;

- Information added by sentiment over message volume when evaluated against asset returns: 2.46%;
- Information added by sentiment over message volume when evaluated against absolute asset returns: 2.41%.

The results of this sensitivity experiment are given in the table below:

	Present study's results (1-hour discretisation window size)		Sensitivity experiment (5-min discretisation window size)	
Instrument and Twitter filter Type ↓	**Maximum information added by sentiment over message volume when evaluated against asset returns:**	**Maximum information added by sentiment over message volume when evaluated against absolute asset returns:**	**Maximum information added by sentiment over message volume when evaluated against asset returns:**	**Maximum information added by sentiment over message volume when evaluated against absolute asset returns:**
Combination A: The Home Depot, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	0.79%	0.58%	4.35%	4.18%
Combination B: J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	2.73%	2.57%	1.07%	1.01%
Combination C: Apple, Inc. CFDs. Ticker-ID only Filter:	2.46%	2.41%	0.63%	0.36%

TABLE 18: SENSITIVITY TO DISCRETISATION WINDOW SIZE: ROBUSTNESS RESULTS FOR MESSAGE SENTIMENT ADDING INFORMATION TO WHAT IS AVAILABLE FROM MESSAGE VOLUMES

Table 18 compares results from experiments using the 1-hour discretisation window size to comparative results under parameter variation (i.e., using the 5-minute discretisation window size), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that message sentiment adds information over what is attainable from message volumes. The following observations are identified:

- For the asset for which a minimal quantity of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:

  - Sentiment continues to add information over message volumes, as is the case with the 1-hour discretisation window size experiments;

  - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the 1-hour discretisation window size experiments;

  - More information is added by sentiment over message volumes when evaluating data discretised to 5-minute windows, than when discretised to 1-hour windows. Therefore, it can be shown that discretisation of the study's dataset to a higher resolution than 1-hour yields a larger amount of information added by message sentiment to message volume for Financial-instrument/Twitter-Filter combinations which experience minimal message information being added by message sentiment over what is attainable from message volumes if discretised to hourly windows.

- For the asset for which a maximal quantity of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:

  - Sentiment continues to add information over message volumes, as is the case with the 1-hour discretisation window size experiments;

  - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the 1-hour discretisation window size experiments;

  - Less information is added by sentiment over message volumes when evaluating data discretised to 5-minute windows, than when discretised to 1-hour windows. Therefore, it can be shown that discretisation of the study's dataset to a higher resolution than 1-hour yields a smaller amount of information added by message sentiment to message volume for Financial-instrument/Twitter-Filter combinations which experience

    - maximal message information being added by message sentiment over what is attainable from message volumes if discretised to hourly windows.

- For the asset for which an intermediate amount of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., Apple, Inc. CFDs from the Ticker-ID only Twitter Filter, results from the 5-minute discretisation window size experiments show that:
    - Sentiment continues to add information over message volumes, as is the case with the 1-hour discretisation window size experiments;
    - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the 1-hour discretisation window size experiments;
    - Less information is added by sentiment over message volumes when evaluating data discretised to 5-minute windows, than when discretised to 1-hour windows.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for message sentiment adding information over what is attainable from message volumes are robust against variation in the discretisation-window size parameter when evaluated against the highest common resolution within the study's dataset. Even at a high-resolution discretisation window size, the study continues to identify that sentiment adds information over what is attainable from message volumes across the range of assets' characteristics within the study's dataset. Furthermore, this robustness experiment has shown that a higher resolution discretisation window size yields an increase in the information added by message sentiment relative to message volume for Financial-instrument/Twitter-Filter combinations where message sentiment adds least information to message volume when discretised to the original hourly window size. In contrast, this robustness experiment has shown that a higher resolution discretisation window size yields a decrease in the information added by message sentiment relative to message volume for Financial-instrument/Twitter-Filter combinations where message sentiment adds most information to message volume when discretised to the original hourly window size. This indicates that whilst the study's results for message sentiment adding information over what is attainable from message volumes are robust against variation in the discretisation-window parameter when evaluated against the highest common window resolution within the study's dataset, additional insight can be extracted from sentiment over what is attainable from message volumes by discretising the data to higher resolutions in instances where minimal additional information can be harnessed over message volumes.

7.1.2 Sensitivity to discretisation window size: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive

The Financial-instrument/Twitter-Filter combinations selected for this sensitivity experiment, according to the criteria in List 3, are detailed below.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 1.1 messages per minute;
    - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 2.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study’s findings as listed in Chapter 6.3.2:
    - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study’s dataset: 184.0 messages per minute;
    - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 14.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study’s findings as listed in Chapter 6.3.2:
    - Of the assets listed in Table 17, eleven used the Ticker-ID AND/OR Company Name Twitter Filter. The mean message volume over the study’s dataset for these eleven assets is: 48.3 messages per minute. The asset which attracted a mean message volume over the study’s dataset closest to this rate is: McDonald’s, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 7.

The results of this sensitivity experiment are given in the table below:

		Present study's results (1-hour discretisation window size)	**Sensitivity experiment (5-min discretisation window size)**
Instrument & Twitter filter Type ↓	**Mean message volume over the study's 3-month dataset:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**
Combination D: J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	1.1	2	24
Combination E: Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	184.0	14	15
Combination F: McDonald's Corp. CFDs. Ticker-ID only Filter:	48.3	7	7

Table 19: Sensitivity to discretisation window size: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive

Table 19 compares the results of experiments using the 1-hour discretisation window size to comparative results under parameter variation (i.e., using the 5-minute discretisation window size), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive. The following observations are identified:

- For the asset for which minimal message volume is observed in the study’s findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:
  - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the 5-minute discretisation window size experiments. However, the number of such time-shifts increases substantially from 2 for the 1-hour discretisation window experiments to 24 for the 5-minute discretisation window experiments.
- For the asset for which maximal message volume is observed in the study’s findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from the 5-minute discretisation window size experiments show that:

    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the 5-minute discretisation window size experiments. The number of such time-shifts remains near-constant, with 14 for the 1-hour discretisation window experiments and 15 for the 5-minute discretisation window experiments.

- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the 5-minute discretisation window size experiments. The number of such time-shifts remains constant, with 7 for the 1-hour discretisation window experiments and 7 for the 5-minute discretisation window experiments.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results of a greater message volume indicating the possibility that social media sentiment is more predictive are robust against variation in the discretisation-window size parameter when evaluated against the highest common window resolution within the study's dataset. The number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns is either identical or greater for experiments when the data are discretised to the high-resolution 5-minute windows, when compared to the study's experiments for which data are discretised to hourly windows.

It is however observed that for the study's entire dataset, as the mean message volume per Financial-instrument/Twitter-Filter combination decreases, the number of statistically-significant leading information surplus time-shifts from sentiment increases when evaluated against asset returns. This therefore highlights that whilst the study's parameters are robust to discretisation window size variation with regards to a greater message volume indicating the possibility that social media sentiment is more predictive, the use of higher-resolution discretisation windows also makes social media sentiment more predictive for Financial-instrument/Twitter-Filter combinations which attract lower message volumes.

7.1.3 Sensitivity to discretisation window size: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.1.2.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 1.1 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 12 hours.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 184.0 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 14 hours.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2:
  - McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
  - Mean message volume per minute over the study's dataset: 48.3 messages per minute;
  - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 13 hours.

The results of this sensitivity experiment are given in the table below:

		Present study's results (1-hour discretisation window size)	**Sensitivity experiment (5-min discretisation window size)**
Instrument & Twitter filter Type ↓	**Mean message volume over the study's 3-month dataset:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**
Combination D: J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	1.1	12 hours	17 hours
Combination E: Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	184.0	14 hours	18 hours
Combination F: McDonald's Corp. CFDs. Ticker-ID only Filter:	48.3	13 hours	14 hours

TABLE 20: SENSITIVITY TO DISCRETISATION WINDOW SIZE: ROBUSTNESS RESULTS FOR A GREATER MESSAGE VOLUME INDICATING THE POSSIBILITY THAT SOCIAL MEDIA SENTIMENT IS MORE PREDICTIVE MORE IN ADVANCE

Table 20 compares the results of experiments using the 1-hour discretisation window size to comparative results under parameter variation (i.e., using the 5-minute discretisation window size), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive more in advance. The following observations are identified:

- For the asset for which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist under the 5-minute discretisation window size experiments. However, the time-shift at which this occurs increases from 12 hours for the 1-hour discretisation window experiments to 17 hours for the 5-minute discretisation window experiments.
- For the asset for which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from the 5-minute discretisation window size experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist under the 5-minute discretisation window size experiments. However, the time-shift at which this occurs increases from 14 hours for the 1-hour discretisation window experiments to 18 hours for the 5-minute discretisation window experiments.
- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the 5-minute discretisation window size experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist under the 5-minute discretisation window size experiments. The time-shift at which this occurs remains near-constant, rising from 13 hours for the 1-hour discretisation window experiments to 14 hours for the 5-minute discretisation window experiments.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against variation in the discretisation-window size parameter when evaluated against the highest common resolution within the study's dataset. The leading time-shift corresponding to the largest information surplus from Twitter sentiment is either identical or greater for experiments when the data are discretised to high-resolution 5-minute windows, when compared to the study's experiments for which data are discretised to hourly windows.

It is however observed that for the study's entire dataset, as the mean message volume per Financial-instrument/Twitter-Filter combination decreases, the time-shift corresponding to the largest information surplus from Twitter sentiment increases. This therefore highlights that whilst the study's parameters are robust to discretisation window size variation with regards to a greater message volume indicating the possibility that social media sentiment is more predictive more in advance, the use of higher-resolution discretisation windows also makes social media sentiment more predictive more in advance for Financial-instrument/Twitter-Filter combinations which attract lower message volumes.

7.1.4 Summary of robustness of the study’s results to variation in discretisation window size

For ease of comparison, the study’s *main* results for the Financial-instrument/Twitter-Filter combinations selected according to the criteria in List 3 are presented below in Table 21. N.B., this table is an extract from the larger Table 17 which contains a summary of the positive results for social media’s ability to lead securities’ returns ahead of time. Consequently, detailed explanations of the “N/A” statements in this table are given in the Results section of this Thesis (Chapter 6). A summary of the results of the robustness experiments relating to variation in discretisation window size for the same Financial-instrument/Twitter-Filter combinations are subsequently presented below in Table 22.

Filter ID	**Instrument**	**Filter type**	**Mean message volume per minute**	**Largest statistically-significant information surplus from sentiment vs. returns**	**Largest statistically-significant information surplus from message volume vs. returns**	**Largest statistically-significant information surplus from message volume vs. absolute returns**
2	Apple, Inc. CFDs	Ticker-ID	1.8	3.35%	0.89%	0.94%
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184	2.64%	N/A	N/A
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	1.9	2.81%	2.02%	2.23%
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	3.94%	1.21%	1.37%
29	McDonald’s, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	1.90%	N/A	N/A

Filter ID	**Instrument**	**Leading time-shift corresponding to the largest information surplus from sentiment vs. returns**	**Sentiment type corresponding to the largest information surplus**	**Number of statistically-significant leading information surplus time-shifts from sentiment vs. returns**	**Number of statistically-significant leading information surplus time-shifts from message volume vs. returns**	**Number of statistically-significant leading information surplus time-shifts from message volume vs. absolute returns**
2	Apple, Inc. CFDs	14 hours	Negative	2	5	9
15	Google, Inc. CFDs	14 hours	Net	14	N/A	N/A
17	The Home Depot, Inc. CFDs	11 hours	Positive	8	4	4
25	J.P. Morgan, Inc. CFDs	12 hours	Positive	2	15	14
29	McDonald’s, Corp. CFDs	13 hours	Net	7	N/A	N/A

TABLE 21: ORIGINAL RESULTS FOR THE FINANCIAL-INSTRUMENT/TWITTER-FILTER COMBINATIONS SELECTED FOR THE TESTS OF THE ROBUSTNESS OF THE STUDY’S RESULTS

In comparison, Table 22 shows the results of the experiments which explore the robustness of the study's findings to variation in discretisation window size, when 5-minute windows are used instead of hourly windows (as explained in Chapter 7.1).

Filter ID	Instrument	Filter type	Mean message volume per minute	Largest statistically-significant information surplus from sentiment vs. returns	Largest statistically-significant information surplus from message volume vs. returns	Largest statistically-significant information surplus from message volume vs. absolute returns
2	Apple, Inc. CFDs	Ticker-ID	1.8	4.30%	3.67%	3.93%
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184	1.81%	0.12%	0.23%
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	1.9	5.85%	1.50%	1.67%
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	2.09%	1.01%	1.07%
29	McDonald's, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	0.98%	0.24%	0.27%

Filter ID	Instrument	Leading time-shift corresponding to the largest information surplus from sentiment vs. returns	Sentiment type corresponding to the largest information surplus	Number of statistically-significant leading information surplus time-shifts from sentiment vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. absolute returns
2	Apple, Inc. CFDs	16 hours	Negative	19	11	12
15	Google, Inc. CFDs	18 hours	Net	15	6	14
17	The Home Depot, Inc. CFDs	12 hours	Negative	43	13	13
25	J.P. Morgan, Inc. CFDs	17 hours	Net	24	14	13
29	McDonald's, Corp. CFDs	14 hours	Positive	7	5	5

TABLE 22: RESULTS OF THE STUDY'S EXPERIMENTS UNDER VARIATION IN DATA DISCRETISATION WINDOW SIZE. RESULTS ARE PRESENTED FOR EXPERIMENTS USING DISCRETISATION WINDOWS OF 5-MINUTES IN SIZE

As detailed in Chapters 7.1.1 to 7.1.3, the study's results were tested for sensitivity to parameter variation relating to discretisation window size. By changing the study's fundamental discretisation window size from 1-hour (as explained in Chapters 5.3 and 5.6.1) to 5-minutes (as explained in Chapter 7.1), the robustness of the study's *main* results from its information theory analysis experiments (as detailed in Chapter 6.3) is established.

It has been shown that for the spectrum of assets considered by this robustness experiment:

- The study's results for message sentiment adding information to what is available from message volumes are robust against variation in the discretisation-window size parameter when evaluated against the highest common resolution within the study's dataset. However, this robustness experiment has shown that additional insight can be extracted from sentiment over message volume by discretising the data to higher resolutions in instances where minimal additional information can be harnessed over message volume.
- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive are robust against variation in the discretisation-window size parameter when evaluated against the highest common resolution within the study's dataset. However, this robustness experiment has shown that the use of higher-resolution discretisation windows also makes social media sentiment more predictive for Financial-instrument/Twitter-Filter combinations which attract lower message volumes.
- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against variation in the discretisation-window size parameter when evaluated against the highest common resolution within the study's dataset. However, this robustness experiment has shown that the use of higher-resolution discretisation windows also makes social media sentiment more predictive more in advance for Financial-instrument/Twitter-Filter combinations which attract lower message volumes.

Whilst the study's results are therefore robust against discretisation-window size changes as per the parameter variations performed in this chapter, it is highlighted that additional insight can be gained from social media sentiment for low message-volume assets if using higher-resolution discretisation windows.

7.2 Sensitivity to discretisation window offset

The study’s results were determined based on the analysis of the dependency between Twitter data and financial data using discretisation windows on the hour (as detailed in Chapter 5.3). It is anticipated that offsetting the financial data and the social media data using windows not on the hour (e.g. such that adjacent discretised windows occur at 30-minutes past each hour) will have minimal effect on the study’s results, when all other conditions in List 1 are kept constant. This is because the dispersion of a random sample of Tweets across the hour is largely homogenous.

Consider Figure 25 below, which shows the mean number of Tweets detected per minute past the hour across the study’s dataset using the broad, topic-unspecific capture of random Tweets from the US (as per Filter IDs 38 or 39 in Table 1).

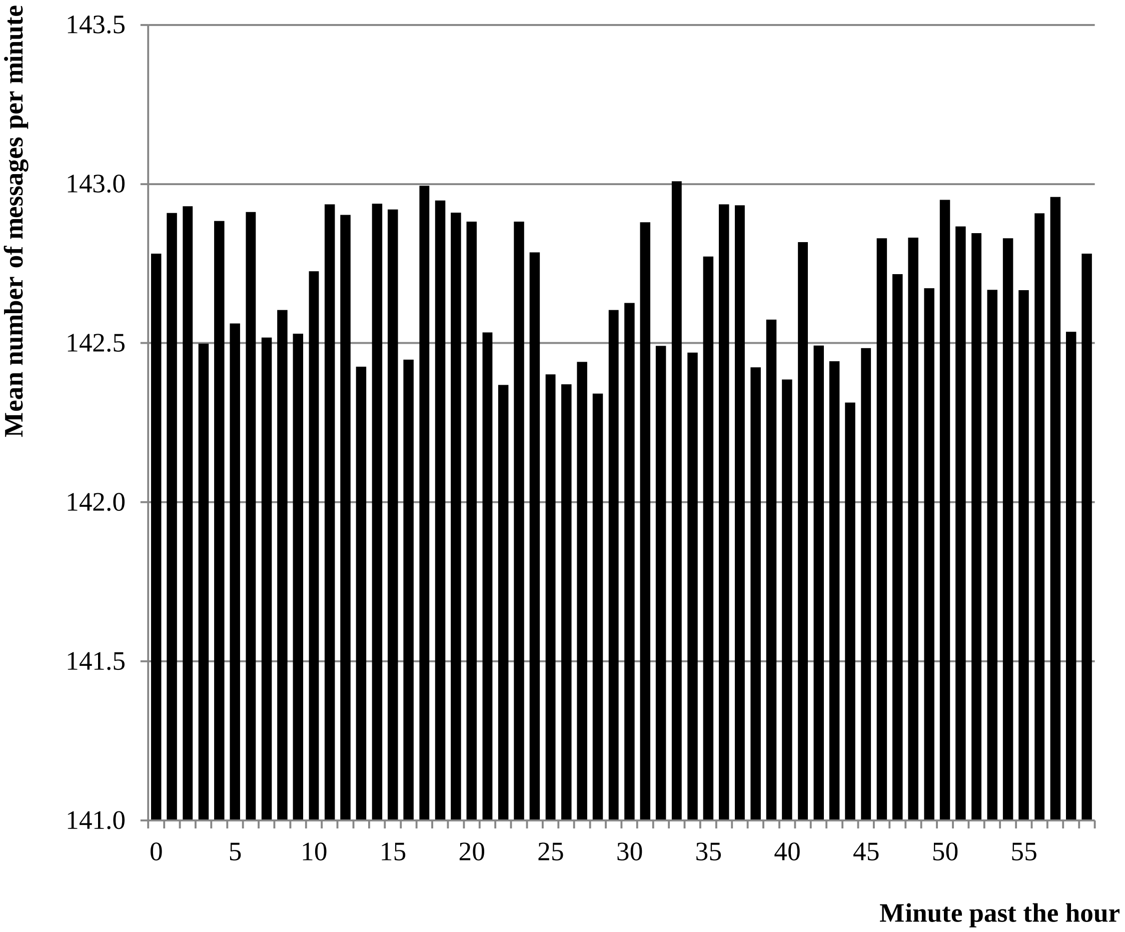


FIGURE 25: MEAN VOLUME OF MESSAGES PER MINUTE PAST THE HOUR FOR STRING-UNFILTERED TWEETS FROM THE US

Figure 25 visualises the mean number of US-Tweet, per minute past the hour across the study's dataset (with a mean of 142.7 and a standard deviation of 0.2109 messages per minute). This low dispersion therefore suggests that off-hour discretisation will have minimal effect on the robustness of the study's results.

To confirm this, what is needed is a repeat of the study's information theory analysis experiments on key Financial-instrument/Twitter-Filter combinations (as detailed in List 3) using a non-hourly offset window between the social media and financial dataset. This non-hourly offset should be a factor of an hour to mirror common financial trading data discretisations, and as a by-product, be straightforward to manipulate during the calculations process.

Thus, to test the sensitivity of the study's results to non-hourly discretisation window offsets, the analysis theory described in Chapter 5.3.2 and implemented via the methodology described in Chapter 5.5 was repeated for a select range of assets according to the criteria in List 3 using a +30-minute discretisation window offset. A +30-minute discretisation window offset is chosen as it offers a balanced encapsulation of the continuous (undiscretised) social media and financial data by splitting the hour into two time-frames of equal (30 minute) width. The effect of this +30-minute offset window is shown diagrammatically below in Figure 26.

Instituting a +30-minute discretisation window offset into the data discretisation process has the effect of offsetting the aggregation windows to which the continuous social media and financial data are appended by way of arithmetic mean averaging (as discussed in Chapter 5.3) into adjacent windows that are not on the hour. In this study, the social media and financial data were discretised into hourly windows occurring on the hour (as denoted by the yellow arrows). To test the robustness of the study’s results, its information theory analysis experiments were repeated on key Financial-instrument/Twitter-Filter combinations (as detailed in List 3) using discretisation windows offset by +30-minutes. This process discretised the social media and financial data such that the aggregation windows occur at 30-minutes past the hour (as denoted by the blue arrows).

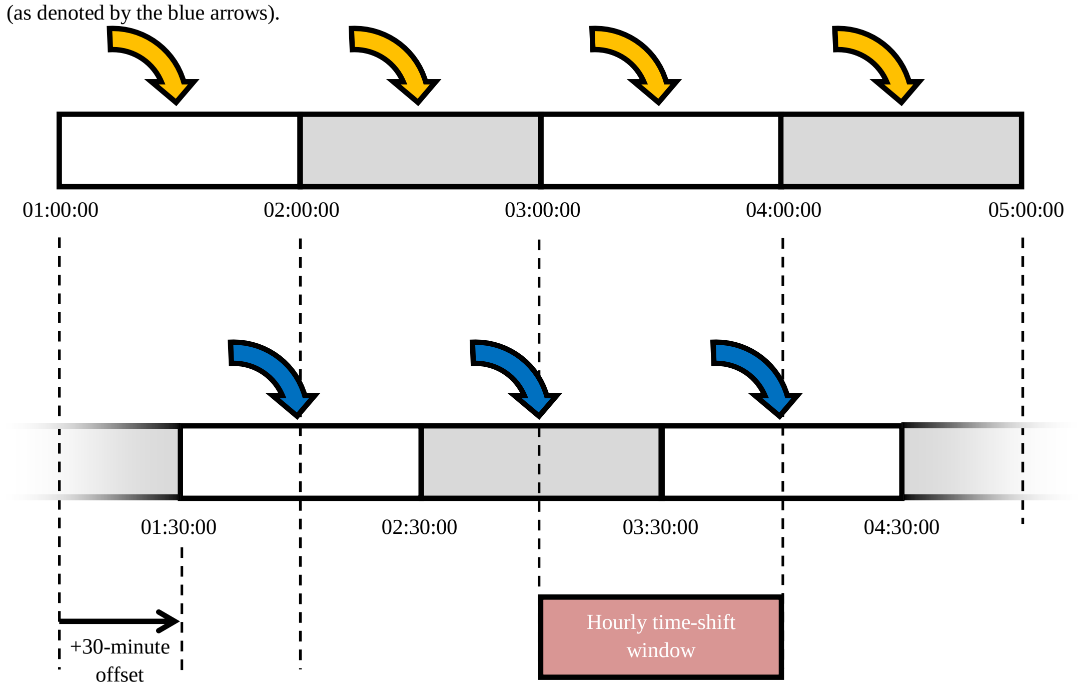


FIGURE 26: EXAMPLE OF THE EFFECT OF IMPLEMENTING A +30-MINUTE OFFSET TO DISCRETISATION WINDOWS

Figure 26 also shows a representative hourly time-shift window. Under the condition of a +30-discretisation window offset, the location of the hourly time-shifts is therefore also offset by +30-minutes. This is a by-product of the procedures behind the implementation of the discretisation window offsets.

The results of this robustness experiment are presented in the subsequent subchapters, and in such manner address the second parameter variation point in List 1.

7.2.1 Sensitivity to discretisation window offset: robustness results for message sentiment adding information over what is attainable from message volumes

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.1.1.

Financial-instrument/Twitter-Filter combination A:

- For which a minimal quantity of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Information added by sentiment over message volume when evaluated against asset returns: 0.79%;
    - Information added by sentiment over message volume when evaluated against absolute asset returns: 0.58%.

Financial-instrument/Twitter-Filter combination B:

- For which maximal information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Information added by sentiment over message volume when evaluated against asset returns: 2.73%;
    - Information added by sentiment over message volume when evaluated against absolute asset returns: 2.57%.

Financial-instrument/Twitter-Filter combination C:

- For which an intermediate amount of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
  - Apple, Inc. CFDs from the Ticker-ID only Twitter Filter;
  - Information added by sentiment over message volume when evaluated against asset returns: 2.46%;
  - Information added by sentiment over message volume when evaluated against absolute asset returns: 2.41%.

The results of this sensitivity experiment are given in the table below:

	Present study's results (On-hour discretisation offset)		**Sensitivity experiment (+30-min discretisation offset)**	
Instrument and Twitter filter ↓	**Maximum Information added by sentiment over message volume when evaluated against asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against absolute asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against absolute asset returns:**
Combination A: **The Home Depot, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	0.79%	0.58%	0.78%	0.58%
Combination B: **J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	2.73%	2.57%	2.71%	2.57%
Combination C: **Apple, Inc. CFDs. Ticker-ID only Filter:**	2.46%	2.41%	2.43%	2.39%

TABLE 23: SENSITIVITY TO DISCRETISATION WINDOW OFFSET: ROBUSTNESS RESULTS FOR MESSAGE SENTIMENT ADDING INFORMATION OVER WHAT IS ATTAINABLE FROM MESSAGE VOLUMES

Table 23 compares the results of experiments using on-hour data discretisation to comparative results under parameter variation (i.e., using a +30-minute discretisation window offset), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that message sentiment adds information over what is attainable from message volumes. The following observations are identified:

- For the asset for which a minimal quantity of information is added by message sentiment over what is attainable from message volumes in the study’s findings as listed in Chapter 6.3.2, i.e., The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - Sentiment continues to add information over message volumes, as is the case with the on-hour discretisation window size experiments;
    - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the on-hour discretisation window size experiments;
    - A similar quantity of information is added by sentiment over message volume when evaluating data discretised using a +30-minute offset, compared to when the data are discretised to on-hour windows. The percentage change between the quantities of information added by sentiment over message volume when using a +30-minute offset is -1.27%, when compared to the quantities of information added by sentiment over message volume when using on-hour discretisation (0.79% to 0.78%, respectively). Therefore, it can be shown that discretisation of the study’s dataset using a +30-minute discretisation window offset data offset has minimal effect on the study’s findings for an asset for which minimal information is added by message sentiment over what is attainable from message volumes.
- For the asset for which a maximal quantity of information is added by message sentiment over what is attainable from message volumes in the study’s findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - Sentiment continues to add information over message volumes, as is the case with the on-hour discretisation window size experiments;

    - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the on-hour discretisation window size experiments;
    - A similar quantity of information is added by sentiment over message volume when evaluating data discretised using a +30-minute offset, compared to when the data are discretised to on-hour windows. The percentage change between the quantities of information added by sentiment over message volume when using a +30-minute offset is -0.73%, when compared to the quantities of information added by sentiment over message volume when using on-hour discretisation (2.73% to 2.71%, respectively). Therefore, it can be shown that discretisation of the study's dataset using a +30-minute discretisation window offset data offset has minimal effect on the study's findings for an asset for which maximal information is added by message sentiment over what is attainable from message volumes.

- For the asset for which an intermediate amount of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., Apple, Inc. CFDs from the Ticker-ID only Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - Sentiment continues to add information over message volumes, as is the case with the on-hour discretisation window size experiments;
    - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with the on-hour discretisation window size experiments;
    - A similar quantity of information is added by sentiment over message volume when evaluating data discretised using a +30-minute offset, compared to when the data are discretised to on-hour windows. The percentage change between the quantities of information added by sentiment over message volume when using a +30-minute offset is -

1.22%, when compared to the quantities of information added by sentiment over message volume when using on-hour discretisation (2.46% to 2.43%, respectively). Therefore, it can be shown that discretisation of the study's dataset using a +30-minute discretisation window offset data offset has minimal effect on the study's findings for an asset for which an intermediate quantity of information is added by message sentiment over what is attainable from message volumes.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for message sentiment adding information to what is available from message volumes are robust against variation in the discretisation window offset parameter when evaluated using a +30-minute offset. Even at +30-minute discretisation window offset, the study continues to identify that sentiment adds information over what is attainable from message volumes across the range of assets' characteristics within the study's dataset. Furthermore, this robustness experiment has shown that a discretisation window offset yields little change in the information added by message sentiment relative to message volume, regardless of quantity of information added by sentiment to message volume when originally discretised to on-hour windows. This indicates that the study's results for message sentiment adding information over what is attainable from message volumes are robust against variation in the discretisation-window offset parameter under the conditions of the robustness experiment.

7.2.2 Sensitivity to discretisation window offset: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.1.2.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
  - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
  - Mean message volume per minute over the study's dataset: 1.1 messages per minute;
  - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 2.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2:
  - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
  - Mean message volume per minute over the study's dataset: 184.0 messages per minute;
  - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 14.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 48.3 messages per minute;
    - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 7.

The results of this sensitivity experiment are given in the table below:

		Present study's results (1-hour discretisation window size)	**Sensitivity experiment (+30-min discretisation offset)**
Instrument & Twitter filter Type↓	**Mean message volume over the study's 3-month dataset:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**
Combination D: **J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	1.1	2	3
Combination E: **Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	184.0	14	14
Combination F: **McDonald's Corp. CFDs. Ticker-ID only Filter:**	48.3	7	7

Table 24: Sensitivity to discretisation window offset: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive

Table 24 compares the results of experiments using on-hour discretisation windows to comparative results under parameter variation (i.e., using a +30-minute discretisation window offset), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive. The following observations are identified:

- For the asset for which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the +30-minute discretisation window offset experiments. The number of such time-shifts remains near-constant, with 2 for on-hour discretisation window experiments, and 3 for the +30-minute discretisation window offset experiments.
- For the asset for which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from the +30-minute discretisation window offset experiments show that
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the +30-minute discretisation window offset experiments. The number of such time-shifts remains constant, with 14 for on-hour discretisation window experiments, and 14 for the +30-minute discretisation window offset experiments.
- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist under the +30-minute discretisation window offset experiments. The number of such time-shifts remains constant, with 7 for on-hour discretisation window experiments, and 7 for the +30-minute discretisation window offset experiments.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for a greater message volume indicating the possibility that social media sentiment is more predictive are robust against variation in the discretisation window offset parameter when evaluated using a +30-minute discretisation window offset. The number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns is either identical or greater for experiments when the data are discretised using a +30-minute discretisation window offset, when compared to the study's experiments for which data are discretised on the hour.

7.2.3 Sensitivity to discretisation window offset: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.2.2.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 1.1 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 12 hours.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 184.0 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 14 hours.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 48.3 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 13 hours.

The results of this sensitivity experiment are given in the table below:

		Present study's results (On-hour discretisation offset)	**Sensitivity experiment (+30-min discretisation window size)**
Instrument & Twitter filter Type ↓	**Mean message volume over the study's 3-month dataset:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**
Combination D: **J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	1.1	12 hours	17 hours
Combination E: **Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	184.0	14 hours	18 hours
Combination F: **McDonald's Corp. CFDs. Ticker-ID only Filter:**	48.3	13 hours	14 hours

TABLE 25: SENSITIVITY TO DISCRETISATION WINDOW OFFSET: ROBUSTNESS RESULTS FOR A GREATER MESSAGE VOLUME INDICATING THE POSSIBILITY THAT SOCIAL MEDIA SENTIMENT IS MORE PREDICTIVE MORE IN ADVANCE

Table 25 compares the results experiments using on-hour discretisation windows to comparative results under parameter variation (i.e., using a +30-minute discretisation window offset), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive more in advance. The following observations are identified:

- For the asset for which minimal message volume is observed in the study’s findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns from sentiment continues to exist under the +30-minute discretisation window offset experiments. The time-shift at which this occurs remains constant at 12 hours for the on-hour discretisation window experiments and 12 hours for the +30-minute discretisation window offset experiments.

- For the asset for which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from the +30-minute discretisation window offset experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist under the +30-minute discretisation window offset experiments. The time-shift at which this occurs is near-constant at 14 hours for the on-hour discretisation window experiments and 13 hours for the +30-minute discretisation window offset experiments.
- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the +30-minute discretisation window offset experiments show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist under the +30-minute discretisation window offset experiments. The time-shift at which this occurs is near-constant at 13 hours for the on-hour discretisation window experiments and 12 hours for the +30-minute discretisation window offset experiments.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against variation in the discretisation-window offset parameter when evaluated against when evaluated against a +30-minute discretisation window offset. The leading time-shift corresponding to the largest information surplus from Twitter sentiment is either identical or near-identical for experiments when the data are discretised using a +30-minute window offset, when compared to the study's experiments for which data are discretised to the hour.

7.2.4 Summary of robustness of the study's results to variation in discretisation window offset

For ease of comparison, the study's *main* results for the Financial-instrument/Twitter-Filter combinations selected according to the criteria in List 3 are presented earlier in Table 21. A summary of the results of the robustness experiments relating to variation in discretisation window offset for the same Financial-instrument/Twitter-Filter combinations are then presented below in Table 26. N.B., ***:** the "N/A" statements in this table are due to message sentiment not being statistically-significant in leading the corresponding assets' returns under the +30-minute discretisation window offset parameter – this mirrors the observations seen with on-hour discretisation window offset, as detailed in Chapter 6: Results.

Filter ID	Instrument	Filter type	Mean message volume per minute	Largest statistically-significant information surplus from sentiment vs. returns	Largest statistically-significant information surplus from message volume vs. returns	Largest statistically-significant information surplus from message volume vs. absolute returns
2	Apple, Inc. CFDs	Ticker-ID	1.8	3.32%	0.89%	0.93%
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184	2.66%	N/A*	N/A*
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	1.9	2.82%	2.04%	2.25%
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	3.91%	1.20%	1.34%
29	McDonald's, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	1.90%	N/A*	N/A*

Filter ID	Instrument	Leading time-shift corresponding to the largest information surplus from sentiment vs. returns	Sentiment type corresponding to the largest information surplus	Number of statistically-significant leading information surplus time-shifts from sentiment vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. absolute returns
2	Apple, Inc. CFDs	14 hours	Negative	2	5	9
15	Google, Inc. CFDs	13 hours	Net	14	N/A*	N/A*
17	The Home Depot, Inc. CFDs	11 hours	Positive	8	4	4
25	J.P. Morgan, Inc. CFDs	12 hours	Positive	3	15	14
29	McDonald's, Corp. CFDs	12 hours	Net	7	N/A*	N/A*

TABLE 26: RESULTS OF THE STUDY'S EXPERIMENTS UNDER VARIATION IN DATA DISCRETISATION WINDOW OFFSET. RESULTS ARE PRESENTED FOR EXPERIMENTS USING DISCRETISATION WINDOW OFFSETS OF +30-MINUTES

As detailed in Chapters 7.2.1 to 7.2.3, the study's results were tested for sensitivity to parameter variation relating to discretisation window offset. By changing the study's fundamental discretisation window offset from on-the-hour (as detailed in Chapter 5.3) to an offset of +30-minutes (as explained in Chapter 7.2), the robustness of the study's *main* results from the information theory analysis experiments (as detailed in Chapter 6.3) is established.

It has been shown that for the spectrum of assets considered by this robustness experiment:

- The study's results for message sentiment adding information over what is attainable from message volumes are robust against variation in the discretisation-window offset parameter when evaluated using a +30-minute offset. This robustness experiment has shown that a discretisation window offset yields little or no change in the information added by message sentiment relative to message volume, regardless of quantity of information added by sentiment to message volume when originally discretised to on-hour windows.

- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive are robust against variation in the discretisation window offset parameter when evaluated using a +30-minute discretisation window offset. This robustness experiment has shown that a discretisation window offset yields little or no change in the number of statistically-significant leading information surplus time-shifts from Twitter sentiment.
- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against variation in the discretisation-window offset parameter when evaluated against when evaluated against a +30-minute discretisation window offset. This robustness experiment has shown that a discretisation window offset yields little or no change in the lead-time corresponding to the largest information surplus from Twitter sentiment.

Whilst the study's results are therefore robust against discretisation window offset changes as per the parameter variations performed in this chapter, it is highlighted that no significant additional insight can be gained from alteration of the discretisation window offset parameter.

7.3 Sensitivity to using an alternative mutual information histogram binning methodology

As discussed in Chapter 5.3.2.1, the computation of entropy, which is necessary as part of the process for calculating mutual information, is based on the probability of the values within the dataset being investigated. The probability distributions were estimated in this study using a three-dimensional histogram (see Chapter 5.3.2 for details). The selection of histogram bin sizes was performed using the Sturges' histogram rule[95], often used as a default tool in statistical packages[96].

This study is not focussed on the comparison of histogram binning methods or the identification of the optimum histogram binning method for the study's dataset in the calculations of mutual information. Rather, it is an exploration of the application of an accepted information theory-based measure of dependency using well-documented mathematical processes for the calculation of entropy. Nonetheless, what is needed is an alternative commonplace process for bin size estimation in order to validate that the study's results hold under variation of the third point in List 1.

There is no single 'best' methodology for histogram binning when estimating probability distributions[113] – instead histogram binning methods can be tailored to a specific dataset. However, this study seeks to show that popular mathematical processes that are untailored to the specific dataset can be used to successfully demonstrate that social media data can lead market returns. Therefore, to validate the robustness of the study's results, what is needed is a frequently-used but dataset-untailored alternative to Sturges' histogram rule. A commonplace alternative[96,114] method to Sturges' histogram rule is the Freedman-Diaconis rule[115], and as with the former, the Freedman-Diaconis rule is also often offered in most statistical software packages[96].

The general equation for the Freedman-Diaconis rule is[115]:

$$\omega = 2 \times \mathrm{IQR}(x) n^{-\frac{1}{3}}$$

Where:

- IQR is the interquartile range of the dataset n. This is a measure of statistical-dispersion of data equal to the difference between the upper and lower quartiles of the dataset;
- n is the number of elements in the dataset x;
- ω is the ideal bin width to be used for the histogram;
- Calculating r/ω gives the number of bins for the dataset.

The analysis theory described in Chapter 5.3.2 and implemented via the methodology described in Chapter 5.5 was repeated on key Financial-instrument/Twitter-Filter combinations (as detailed in List 3) using the Freedman-Diaconis rule instead of Sturges' histogram rule for histogram binning in the calculations of entropy. This is based on the observation that the two methodologies are both well-documented and commonplace within statistical-software packages. The results of this robustness experiment are presented in the subsequent subchapters, and in such manner address the third parameter variation point in List 1.

7.3.1 Sensitivity to using an alternative mutual information histogram binning methodology: robustness results for message sentiment adding information over what is attainable from message volumes

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.1.1.

Financial-instrument/Twitter-Filter combination A:

- For which minimal information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Information added by sentiment over message volume when evaluated against asset returns: 0.79%;
    - Information added by sentiment over message volume when evaluated against absolute asset returns: 0.58%.

Financial-instrument/Twitter-Filter combination B:

- For which maximal information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Information added by sentiment over message volume when evaluated against asset returns: 2.73%;
    - Information added by sentiment over message volume when evaluated against absolute asset returns: 2.57%.

Financial-instrument/Twitter-Filter combination C:

- For which an intermediate amount of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2:
    - Apple, Inc. CFDs from the Ticker-ID only Twitter Filter;
    - Information added by sentiment over message volume when evaluated against asset returns: 2.46%;
    - Information added by sentiment over message volume when evaluated against absolute asset returns: 2.41%.

The results of this sensitivity experiment are given in the table below:

	Present study's results (Sturges' Rule histogram binning method)		Sensitivity experiment (Freedman-Diaconis rule histogram binning method)	
Instrument and Twitter filter Type ↓	**Maximum Information added by sentiment over message volume when evaluated against asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against absolute asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against asset returns:**	**Maximum Information added by sentiment over message volume when evaluated against absolute asset returns:**
Combination A: **The Home Depot, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	0.79%	0.58%	0.63%	0.62%
Combination B: **J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	2.73%	2.57%	2.70%	2.76%
Combination C: **Apple, Inc. CFDs. Ticker-ID only Filter:**	2.46%	2.41%	1.50%	1.59%

TABLE 27: SENSITIVITY TO USING AN ALTERNATIVE MUTUAL INFORMATION HISTOGRAM BINNING METHODOLOGY: ROBUSTNESS RESULTS FOR MESSAGE SENTIMENT ADDING INFORMATION OVER WHAT IS ATTAINABLE FROM MESSAGE VOLUMES

Table 18 compares the results of experiments using Sturges' rule for histogram binning to comparative results under parameter variation (i.e., the Freedman-Diaconis rule for histogram binning), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that message sentiment adds information over what is attainable from message volumes. The following observations are identified:

- For the asset for which minimal information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., The Home Depot, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the Freedman-Diaconis histogram binning experiments show that:
  - Sentiment continues to add information over message volumes, as is the case with using Sturges' rule for histogram binning;
  - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with using Sturges' rule for histogram binning;

  - A decrease is seen in the information added by sentiment over message volumes when evaluating the data using the Freedman-Diaconis rule for histogram binning, compared to when evaluating the data using Sturges' rule for histogram binning. The percentage change between the quantities of information added by sentiment over message volume when using the Freedman-Diaconis rule for histogram binning is –20.25%, when compared to the quantities of information added by sentiment over message volume when using Sturges' rule for histogram binning (0.79% to 0.63%, respectively). Therefore it can be shown that the use of the Freedman-Diaconis rule for histogram binning has a strong effect on the study's findings for an asset for which a low quantity of information is added by message sentiment over what is attainable from message volumes.

- For the asset for which maximal information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from the Freedman-Diaconis histogram binning experiments show that:

  - Sentiment continues to add information over message volumes, as is the case with using Sturges' rule for histogram binning;

  - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with using Sturges' rule for histogram binning;

  - A similar quantity of information is added by sentiment over message volumes when evaluating the data using the Freedman-Diaconis rule for histogram binning, compared to when evaluating the data using Sturges' rule for histogram binning. The percentage change between the quantities of information added by sentiment over message volume when using the Freedman-Diaconis rule for histogram binning is –1.10%, when compared to the quantities of information added by sentiment over message volume when using Sturges' rule for histogram binning (2.73% to 2.70%, respectively). Therefore it can be shown that the use of the

    Freedman-Diaconis rule for histogram binning has minimal effect on the study's findings for an asset for which minimal information is added by message sentiment over what is attainable from message volumes.

- For the asset for which an intermediate quantity of information is added by message sentiment over what is attainable from message volumes in the study's findings as listed in Chapter 6.3.2, i.e., Apple, Inc. CFDs from the Ticker-ID only Twitter Filter, results from the Freedman-Diaconis histogram binning experiments show that:

    - Sentiment continues to add information over message volumes, as is the case with using Sturges' rule for histogram binning;

    - More information is provided by the evaluation of message volumes against absolute asset returns, in comparison to evaluation of message volumes against actual asset returns, as is the case with using Sturges' rule for histogram binning;

    - A decrease is seen in the information added by sentiment over message volumes when evaluating the data using the Freedman-Diaconis rule for histogram binning, compared to when evaluating the data using Sturges' rule for histogram binning. The percentage change between the quantities of information added by sentiment over message volume when using the Freedman-Diaconis rule for histogram binning is –39.02%, when compared to the quantities of information added by sentiment over message volume when using Sturges' rule for histogram binning (2.46% to 1.50%, respectively). Therefore it can be shown that the use of the Freedman-Diaconis rule for histogram binning has a strong effect on the study's findings for an asset for which an intermediate quantity of information is added by message sentiment over what is attainable from message volumes.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results are robust against using the Freedman-Diaconis rule for histogram binning for all considered assets: when using this commonplace alternative to Sturges' rule, sentiment continues to add information over what is attainable from message volumes. However, the use of the Freedman-Diaconis rule can diminish the amount of information added by sentiment over what is attainable from message volumes in cases where low quantities of information are added by message sentiment over what is attainable from message volume when using Sturges' rule for histogram binning.

7.3.2 Sensitivity to using an alternative mutual information histogram binning methodology: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.1.2.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 1.1 messages per minute;
    - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 2.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 184.0 messages per minute;

- Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 14.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2:
  - McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
  - Mean message volume per minute over the study's dataset: 48.3 messages per minute;
  - Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns: 7.

The results of this sensitivity experiment are given in the table below:

		Present study's results (Sturges' Rule histogram binning method)	**Sensitivity experiment (Freedman-Diaconis rule histogram binning method)**
Instrument & Twitter filter Type ↓	**Mean message volume over the study's 3-month dataset:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**	**Number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns:**
Combination D: J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	1.1	2	3
Combination E: Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:	184.0	14	19
Combination F: McDonald's Corp. CFDs. Ticker-ID only Filter:	48.3	7	6

TABLE 28: SENSITIVITY TO USING AN ALTERNATIVE MUTUAL INFORMATION HISTOGRAM BINNING METHODOLOGY: ROBUSTNESS RESULTS FOR A GREATER MESSAGE VOLUME INDICATING THE POSSIBILITY THAT SOCIAL MEDIA SENTIMENT IS MORE PREDICTIVE

Table 28 compares the results of experiments using Sturges' rule for histogram binning to comparative results under parameter variation (i.e., the Freedman-Diaconis rule for histogram binning), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive. The following observations are identified:

- For the asset for which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist when using the Freedman-Diaconis rule for histogram binning. The number of such time-shifts remains near-constant, with 2 for experiments using Sturges' rule for histogram binning, and 3 when using the Freedman-Diaconis rule for histogram binning.
- For the asset for which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist when using the Freedman-Diaconis rule for histogram binning. The number of such time-shifts increases, with 14 for experiments using Sturges' rule for histogram binning, and 19 when using the Freedman-Diaconis rule for histogram binning.
- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
    - Leading information surplus time-shifts from sentiment when evaluated against asset returns continue to exist when using the Freedman-Diaconis rule for histogram binning. The number of such time-shifts is near-

constant, with 7 for experiments using Sturges' rule for histogram binning, and 6 when using the Freedman-Diaconis rule for histogram binning.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for a greater message volume indicating the possibility that social media sentiment is more predictive are robust against using a popular alternative method to the Sturges' rule for histogram binning (by instead using the Freedman-Diaconis rule). For low or medium mean message volume assets, the number of statistically-significant leading information surplus time-shifts from sentiment when evaluated against asset returns is either identical or near-identical for experiments when using the Freedman-Diaconis rule for histogram binning, when compared to the study's experiments which use Sturges' rule for histogram binning. However, for assets for which maximal message volume is observed in the study's findings, the use of the Freedman-Diaconis rule for histogram binning results in a larger number of statistically-significant information surplus time-shifts from sentiment when evaluated against asset returns.

7.3.3 Sensitivity to using an alternative mutual information histogram binning methodology: robustness results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance

The criteria for asset selection for this test of the robustness of the study's results are the same as for Chapter 7.3.2.

Financial-instrument/Twitter-Filter combination D:

- For which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2:
    - J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study's dataset: 1.1 messages per minute;

    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 12 hours.

Financial-instrument/Twitter-Filter combination E:

- For which maximal message volume is observed in the study’s findings as listed in Chapter 6.3.2:
    - Google, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study’s dataset: 184.0 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 14 hours.

Financial-instrument/Twitter-Filter combination F:

- For which a mean message volume is observed in the study’s findings as listed in Chapter 6.3.2:
    - McDonald’s, Corp. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter;
    - Mean message volume per minute over the study’s dataset: 48.3 messages per minute;
    - Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns: 13 hours.

The results of this sensitivity experiment are given in the table below:

		Present study's results (Sturges' Rule histogram binning method)	Sensitivity experiment (Freedman-Diaconis rule histogram binning method)
Instrument & Twitter filter Type ↓	**Mean message volume over the study's 3-month dataset:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**	**Leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns:**
Combination D: **J. P. Morgan, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	1.1	12 hours	18 hours
Combination E: **Google, Inc. CFDs. Ticker-ID AND/OR Company Name Filter:**	184.0	14 hours	11 hours
Combination F: **McDonald's Corp. CFDs. Ticker-ID only Filter:**	48.3	13 hours	14 hours

TABLE 29: SENSITIVITY TO USING AN ALTERNATIVE MUTUAL INFORMATION HISTOGRAM BINNING METHODOLOGY: ROBUSTNESS RESULTS FOR A GREATER MESSAGE VOLUME INDICATING THE POSSIBILITY THAT SOCIAL MEDIA SENTIMENT IS MORE PREDICTIVE MORE IN ADVANCE

Table 29 compares the results of experiments using Sturges' rule for histogram binning to comparative results under parameter variation (i.e., the Freedman-Diaconis rule for histogram binning), for the Financial-instrument/Twitter-Filter combinations which are of relevance to this robustness experiment (as detailed in List 3). This is in relation to the generalisation that a greater message volume indicates the possibility that social media sentiment is more predictive more in advance. The following observations are identified:

- For the asset for which minimal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., J.P. Morgan, Inc. CFDs from the Ticker-ID AND/OR Company Name Twitter Filter, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
  - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist when using the Freedman-Diaconis rule for histogram binning. The time-shift at which this occurs increases to 18 hours for experiments using Freedman-Diaconis rule for histogram binning, from 12 hours for experiments using the Sturges' rule for histogram binning.

- For the asset for which maximal message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., Google, Inc. CFDs from the Ticker-ID AND/OR Company Name, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist when using the Freedman-Diaconis rule for histogram binning. The time-shift at which this occurs decreases to 11 hours for experiments using Freedman-Diaconis rule for histogram binning, from 14 hours for experiments using the Sturges' rule for histogram binning.
- For the asset for which a mean message volume is observed in the study's findings as listed in Chapter 6.3.2, i.e., McDonald's, Corp. CFDs from the Ticker-ID AND/OR Company Name, results from experiments using the Freedman-Diaconis rule for histogram binning show that:
    - A leading time-shift corresponding to the largest information surplus from Twitter sentiment when evaluated against asset returns continues to exist when using the Freedman-Diaconis rule for histogram binning. The time-shift at which this occurs remains near-constant, at 13 hours for experiments using Freedman-Diaconis rule for histogram binning, in comparison to 14 hours for experiments using the Sturges' rule for histogram binning.

Therefore, given the spectrum of assets considered by this robustness experiment (the criteria for which are identified in List 3), the study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against using a popular alternative method to the Sturges' rule for histogram binning (by instead using the Freedman-Diaconis rule). In addition, it has been observed that for Financial-instrument/Twitter-Filter combinations which attract a low mean message volume over the study's dataset, the time-shift at which the largest information surplus from Twitter sentiment is detected is more in advance of time for experiments using the Freedman-Diaconis rule for histogram binning when compared to experiments using Sturges' rule for histogram binning. Conversely, it has been observed that for Financial-instrument/Twitter-Filter combinations which attract a high mean message volume over the study's dataset, the time-shift at which the largest information surplus from Twitter sentiment is detected is less in advance of time for experiments using the Freedman-Diaconis rule for histogram binning when compared to experiments using Sturges' rule for histogram binning. This therefore highlights that whilst the study's parameters are robust against histogram binning method changes with regards to a greater message volume indicating the possibility that social media sentiment is more predictive more in advance, the Freedman-Diaconis rule for histogram binning is more suited to low message volume Financial-instrument/Twitter-Filter combinations if seeking to utilise message sentiment more in advance of time.

7.3.4 Summary of robustness of the study's results using an alternative mutual information histogram binning methodology

For ease of comparison, the study's *main* results for the Financial-instrument/Twitter-Filter combinations selected for according to the criteria in List 3 are presented earlier in Table 21. A summary of the results of the robustness experiments relating to using the Freedman-Diaconis rule as an alternative histogram binning method for the same Financial-instrument/Twitter-Filter combinations are then presented below in Table 30.

Filter ID	Instrument	Filter type	Mean message volume per minute	Largest statistically-significant information surplus from sentiment vs. returns	Largest statistically-significant information surplus from message volume vs. returns	Largest statistically-significant information surplus from message volume vs. absolute returns
2	Apple, Inc. CFDs	Ticker-ID	1.8	3.84%	2.34%	2.45%
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	184	6.21%	1.88%	1.91%
17	The Home Depot, Inc. CFDs	Ticker-ID AND/OR Company Name	1.9	2.90%	2.27%	2.28%
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	1.1	3.81%	1.11%	1.15%
29	McDonald's, Corp. CFDs	Ticker-ID AND/OR Company Name	46.5	2.65%	2.08%	2.22%

Filter ID	Instrument	Leading time-shift corresponding to the largest information surplus from sentiment vs. returns	Sentiment type corresponding to the largest information surplus	Number of statistically-significant leading information surplus time-shifts from sentiment vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. returns	Number of statistically-significant leading information surplus time-shifts from message volume vs. absolute returns
2	Apple, Inc. CFDs	16 hours	Net	8	2	2
15	Google, Inc. CFDs	11 hours	Net	19	2	3
17	The Home Depot, Inc. CFDs	12 hours	Positive	7	2	2
25	J.P. Morgan, Inc. CFDs	18 hours	Negative	3	7	6
29	McDonald's, Corp. CFDs	14 hours	Positive	6	4	4

TABLE 30: RESULTS OF THE STUDY'S EXPERIMENTS WHEN USING AN ALTERNATIVE MUTUAL INFORMATION HISTOGRAM BINNING METHODOLOGY (FREEDMAN-DIACONIS RULE INSTEAD OF STURGES' RULE)

As detailed in Chapters 7.2.1 to 7.3.3, the study's results were tested for sensitivity to using an alternative method of histogram binning. As discussed in Chapter 7.3, this test was not focussed on performing an all-encompassing comparison of a range of histogram binning methods. Instead, a popular alternative to the study's primary histogram binning method is selected, and the robustness of the study's *main* results from the information theory analysis experiments (as detailed in Chapter 6.3) is established. It is important to note that in these tests of robustness, as with Sturges' rule for histogram binning, the choice of using the Freedman-Diaconis rule for histogram binning was not based on specifically tailoring it to the study's dataset. Rather, the robustness experiments used this well-documented mathematical process, and popular alternative to the Sturges' rule for determining bin numbers in the calculation of entropy to show that social media data can lead financial data using commonly-available measures of dependency without tailoring it specifically to the study's dataset.

It has been shown that for the spectrum of assets considered by this robustness experiment:

- The study's results for message sentiment adding information over what is attainable from message volumes are robust against a variation in the histogram binning method for all considered assets. However, the use of the Freedman-Diaconis rule for histogram binning can diminish the information added by sentiment to message volume, especially in instances where the amount of information added by sentiment to message-volume is low under the study's original parameters of using Sturges' rule for histogram binning.
- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive are robust against variation in the method for histogram binning when evaluated by using the Freedman-Diaconis rule for histogram binning. In addition, it is shown that using this histogram binning method allows high message-volume assets to be more predictive.
- The study's results for a greater message volume indicating the possibility that social media sentiment is more predictive more in advance are robust against variation in the method for histogram binning when evaluated by using the Freedman-Diaconis rule for histogram binning. In addition, it is shown that using this histogram binning method allows low message-volume assets to be more predictive more in advance.

Whilst the study's results are therefore robust against mutual information histogram binning parameter variations as performed in this chapter, it is highlighted that the use of an alternative but similarly commonplace method for histogram binning can: diminish the additional insight available from message sentiment for low message-volume assets; allow high message-volume assets to be more predictive; and allow low message-volume assets to be more predictive more in advance.

7.4 The 24-hour data-smoothing parameter cannot be varied

As discussed in Chapter 5.6.2.1, under the constraints this study, cyclical patterns were detected in both Twitter sentiments and Twitter volumes. The removal of these cyclical patterns using a backwards-looking Simple Moving Average method provided a clearer view of the true underlying behaviour of the time-series by eliminating the randomness in the data and leaving a smoothed trend-cycle component.

The implementation of a Simple Moving Average required the identification of the number of elements for the window size: too many, and the data would be over-smoothed; too few, and the data would be under-smoothed. To address this issue, the autocorrelation technique was employed, which allows for estimation of the dominating frequency within the social media time-series[102]. Autocorrelation is a representation of the amount of similarity of an observation within a time-series, and another observation within the same time-series, as a function of time separation between such observations[103]. The estimation of the dominating frequency of a discrete signal can be performed by the identification of the largest peak in the autocorrelation function of a time-series occurring at a non-zero lag[104] – by definition, the signal is at its peak autocorrelation at a lag of zero[105]. Compared to the use of the Fourier transform[a106], this methodology is more accurate since the resolution is not limited by the number of samples considered[107].

The study's dataset, under its constraints, demonstrated non-zero peak autocorrelations across the board at lags of 24-hours, an example of which is shown in Figure 14. For this reason, under the methods of the autocorrelation technique the number of elements in the backwards-looking Simple Moving Average calculations had to be fixed at 24.

The variation of smoothing window size cannot be justified mathematically based on the analysis of the study's dataset – this includes the full release of the smoothing window size to zero hours. Its release, or variation, would be contrary to the methods of the autocorrelation technique in the identification of the most-appropriate window size for the smoothing of the randomness within the data, in order to identify the true underlying behaviour of the social media time-series. For this reason, variation in the smoothing-window size parameter is not performed as a test for the robustness of the study's results (fourth point in List 1).

[a] This is a mathematical transformation employed for conversion of signals between time domains and frequency domains.

7.5 Robustness of results summary and derived generalisations

Chapter 7 explores the extent to which the variation in the study's constraints (as detailed in List 1) affects the study's generalised results (as detailed in List 2) according to the possible outcomes of this exercise (as detailed in List 3). The following of the study's constraints were varied independently of one another via a set of segregated experiments:

- Discretisation window size. These experiments explored whether the allocation of the study's dataset to discretisation windows of a higher resolution than 1-hour would produce significant results variation. Given that the financial dataset used in the TCF (see Chapter 5.6.1) had a common highest resolution of 5-minutes, the sensitivity of the study's information theory analysis results were tested by repeating the study's information theory analysis experiments using this highest-resolution discretisation window size.

- Discretisation window offset. These experiments explored the sensitivity of the study's results to non-hourly discretisation window offsets, via implementation of the study's methodology as described in Chapter 5.5 using a +30-minute discretisation window offset. A +30-discretisation window offset was chosen as it offers a balanced encapsulation of the continuous (undiscretised) social media and financial data by splitting the hour into two time-frames of equal (30-minute) width.

- Mutual information histogram binning methodology. Here, the Freedman-Diaconis rule was used instead of Sturges' rule for histogram binning in the calculations of entropy to ascertain if the study's results held when using an alternative but similarly commonplace method of histogram binning. Note that this study is not focussed on the comparison of histogram binning methods or the identification of the optimum histogram binning method for the study's dataset in the calculations of mutual information. Rather, it is an exploration of the application of an accepted information theory-based measure of dependency using well-documented mathematical processes for the calculation of entropy.

Note that justification of why the 24-hour data-smoothing parameter was not varied is given in Chapter 7.4.

Each of the aforementioned tests for robustness was performed to explore the effect on generalisations established from the study's results (as listed in Chapter 6.3.3). These generalisations are that:

1. Message sentiment adds information to what is available from message volumes;
2. A greater message volume indicated the possibility that social media sentiment is more predictive for individual companies;
3. A greater message volume indicated the possibility that social media sentiment is predictive more in advance (i.e., further ahead of time) for individual companies.

In order to perform the tests for robustness, and to explore their effects on the broad aforementioned generalisations, experiments were conducted on key Financial-instrument/Twitter-Filter combinations which are representative of the range of generalised observations seen in Chapter 6.3.3. These key assets consisted of:

- Three assets in total, for which a minimal, maximal and intermediate quantities of information are added by message sentiment over what is attainable from message volumes, respectively, in the study's findings as listed in Chapter 6.3.2;
- Three assets in total, for which minimal, maximal and mean message volume is observed, respectively, in the study's findings as listed in Chapter 6.3.2.

The results of tests for robustness indicate that the study's results are robust against the parameter variations explored throughout Chapter 7 for the key Financial-instrument/Twitter-Filter combinations which are representative of the range of generalised observations seen in Chapter 6.3.3. In particular:

- The study's results are robust against discretisation-window size variation, however additional insight can be gained from social media sentiment for low message-volume assets if using higher-resolution discretisation windows;

- The study's results are robust against discretisation window offset, and no significant additional insight can be gained from alteration of the discretisation window offset parameter;

- The study's results are robust against using an alternative but similarly commonplace method for histogram binning in the calculations of mutual information. However, it is also shown that the use of the Freedman-Diaconis rule for histogram binning instead of Sturges' rule can: diminish the additional insight available from message sentiment for low message-volume assets; allow high message-volume assets to be more predictive; and allow low message-volume assets to be more predictive more in advance;

- Message sentiment continues to add information over what is attainable from message volumes, as shown in Chapters 7.1.1, 7.2.1 and 7.3.1;

- A greater message volume continues to indicate the possibility that social media sentiment is more predictive for individual companies, as shown in Chapters 7.1.2, 7.2.2 and 7.3.2;

- A greater message volume continues to indicate the possibility that social media sentiment is predictive more in advance (i.e., further ahead of time) for individual companies, as shown in Chapters 7.1.3, 7.2.3 and 7.3.3.

LIST 4: SUMMARY OF THE ROBUSTNESS OF THE STUDY'S RESULTS TO PARAMETER VARIATION

The six key Financial-instrument/Twitter-Filter combinations discussed throughout this chapter were used as representative samples, as determined from the criteria in List 3, to test the robustness of the study's results against parameter variation. Given that the study's results have been found to be robust for the six representative and broad-scope Financial-instrument/Twitter-Filter combinations, it can be stated that the rest of the Financial-instrument/Twitter-Filter combinations from the study's results (as summarised in Table 17) are likely to also exhibit the robustness characteristics (as summarised in List 4). This is because the six key Financial-instrument/Twitter-Filter combinations selected for the parameter variation experiments encompass the spectrum of data characteristics of the assets in the study's results, as selected due to the criteria in List 3.

8 SUMMARY AND DISCUSSION

This chapter succinctly summarises the study. An overview of the results and their robustness is presented. The results are then assessed relative to initial hypotheses as listed in Chapter 1.3. Explanations for the study's findings are given. Limitations in the study's methodology are then provided. Finally, a comparison is given to recent works in this research space.

This study centres on the evaluation of string and/or geographically-filtered messages (or 'Tweets') from the Twitter network, to ascertain their ability to lead the returns of market-traded financial securities without any biases associated with profit-maximisation or the implementation of a trading strategy. 112,628,180 Tweets were collected over period from 11th December 2012 to 12th March 2013, and evaluated against the hourly returns of CFDs of US-based publically traded companies; the CFDs/Futures of two popular currency pairs; and the hourly returns of CFDs/Futures of the S&P500 Index and the FTSE100 Index.

The analysis consisted of the evaluation of Tweet sentiments (classification of the polarity of text strings with regards to emotional scales) as well as Tweet message volumes. The sentiment classification was performed using SentiStrength, a package specifically designed for the accurate classification of the short informal text style used in social media.

The study involved the following steps:

1. Filtering and collection of Tweets (Chapters 4.1 and 4.2);
2. Sentiment analysis of Tweets (Chapter 5.1);
3. Exploring measures of dependency suitable for the evaluation of the relationships between Tweets and the returns of market-traded securities (Chapter 5.3);
4. Isolation of the underlying trend component from the Twitter data (Chapter 5.6.2.1);
5. Creation of a metric for measuring the extent to which Twitter data contains lead-time information about the financial data (Chapters 5.5.2 and 5.5.4);

6. Measuring the dependencies between Tweets and the returns of market-traded securities (Chapter 5.5), and testing the results for statistical significance (Chapter 5.5.3);
7. Testing the robustness of the study's results against variations in key parameters (Chapter 7).

The results of the study indicate that when evaluated using an information theory analysis-based measure of dependency (Chapter 5.5), social media contains statistically-significant lead-time information about the returns of market-traded instruments for a limited set of assets (Chapter 6.3). Linear regression analysis-based measures of dependency (Chapter 5.3.1) did not show statistically-significant relationships between Twitter data and the returns of market-traded instruments (Chapter 6.2) – reasons for this are explained in Chapter 8.2.1.

Of the forty-four financial-instrument/Twitter-Filter combinations initially considered (listed in Table 1), the hourly changes in the sentiments of Tweets from twelve Twitter-Filters showed the ability to lead the assets' hourly returns in a statistically-significant manner to a 99% level of confidence. For these twelve financial-instrument/Twitter-Filter combinations, hourly changes in Twitter message sentiments showed a greater ability to lead these assets' hourly returns than hourly changes in Twitter message volumes. The amount of information available from message sentiments is therefore greater than what is available from message volumes. Reasons for this are explained in Chapter 8.2.2.

The study's results were tested for robustness against variation in key parameters (Chapter 7), finding that they are robust against: discretisation-window size variation; the use of an alternative commonplace mutual information histogram method; and discretisation window offset. As a by-product it has also been established that:

- Additional insight can be gained from social media sentiment for low message-volume assets if using higher-resolution discretisation windows;
- No significant additional insight can be gained from alteration of the discretisation window offset parameter;

- Additional insight can be gained from social media by using the Freedman-Diaconis rule in the calculations of mutual information between Twitter data and financial data.

8.1 Assessment of the study's results relative to its hypotheses

To reiterate Chapter 1.3, two hypotheses were explored in this study:

> Hypothesis One: "The analysis of randomised samples of 10% of all Tweets from the United States and the United Kingdom can be used to lead the returns of S&P500 and FTSE100 indices, respectively".
>
> Hypothesis Two: "The analysis of Tweets filtered by instrument identifiers and/or company names can be used to lead the returns of market-traded securities".

The following subchapters are an assessment of the extent to which the aforementioned hypotheses were supported by the study's results.

8.1.1 Hypothesis One

The results of the study show that for the dataset evaluated, large random samples of Tweets from the United States were able to lead the returns of the S&P500 Index. Chapter 6.3.2.1 demonstrates that hourly changes in the sentiments of string-unfiltered Tweets from the US are able to lead hourly changes in the price of S&P500 Index Futures. This is achieved with the study's information theory analysis experiments (as detailed in Chapter 5.5). Here, one time-shift was identified as being leading, with an information surplus of 2.46% at the 99% level of significance occurring at a time-shift of 22-hours achieved via the evaluation of hourly changes in net sentiments. In contrast, the study's linear regression analysis methodology did not show instances of social media leading financial data (as detailed Chapter 6.2). Chapter 8.2.1 discusses why the information theory experiments were able to lead the financial data, whilst the linear regression analysis experiments were not.

The results of the study also show that for the dataset explored, there were no identifiable statistically-significant instances of hourly changes in US string-unfiltered Twitter message sentiments or volumes being able to lead the returns of S&P500 Index CFDs (as discussed in Chapter 6.3.1.6). Similarly, the study was also unable to identify statistically-significant instances of hourly changes in UK string-unfiltered Twitter message sentiments or volumes being able to lead the returns of FTSE100 Index CFDs or Futures (as discussed in Chapters 6.3.1.4 and 6.3.1.5).

The conclusion with regards to Hypothesis One is that for the majority of instances, random samples of Tweets from the United States and the United Kingdom Tweets are not able to lead the returns of those countries' primary stock indices. However, the results do show that in this case of leading the returns of S&P500 Index Futures with string-unfiltered Tweets from the US, Tweet message sentiments do add information over what is attainable from message volumes at the 99% level of significance in the case of one time-shift. The findings of experiments in leading the returns of indices therefore only partially validate Hypothesis One.

8.1.2 Hypothesis Two

The study shows that that for the dataset evaluated, information theory analysis reveals numerous statistically-significant time-shifts at which Tweets filtered by instrument identifiers and/or company names led the returns of market-traded assets. In contrast, the linear regression analysis experiments showed no abilities to lead market data (as detailed Chapter 6.2). Chapter 8.2.1 explains why the information theory experiments were able to lead the financial data, whilst the linear regression analysis experiments were not.

Of the forty-four financial-instrument/Twitter-Filter combinations originally considered by the study:

- Twenty-three are deemed inadmissible because they attract mean message volumes which do not meet the minimum message volume criteria set out in Chapter 6.1. A further two not admitted due to attracting irrelevant messages because of the shortness of the string-filters used. These are listed in Table 5.

- As discussed in Chapter 6.3.1, a further six financial-instrument/Twitter-Filter combinations are rejected because hourly changes in neither the Twitter sentiments nor the Twitter message volumes were able to lead assets' hourly returns.

- This leaves twelve financial-instrument/Twitter-Filter string-filtered combinations for which Twitter data led individual assets' returns to the 99% level of statistical significance. These results are presented in Chapter 6.3.2 and are summarised below in Table 31.

Filter ID	Instrument	Filter type	Do hourly changes in Twitter Message sentiments lead asset returns?	Do hourly changes in Twitter Message volumes lead asset returns?
1	Apple, Inc. CFDs	Ticker-ID AND/OR Company Name	Yes	
2	Apple, Inc. CFDs	Ticker-ID	Yes	Yes
3	Amazon.com, Inc. CFDs	Ticker-ID AND/OR Company Name	Yes	
6	Bank of America, Corp.	Ticker-ID AND/OR Company Name		Yes
8	Cisco Systems, Inc. CFDs	Ticker-ID AND/OR Company Name	Yes	
15	Google, Inc. CFDs	Ticker-ID AND/OR Company Name	Yes	
17	The Home Depot, Inc.	Ticker-ID AND/OR Company Name	Yes	Yes
22	Intel, Corp. CFDs	Ticker-ID AND/OR Company Name	Yes	Yes
25	J.P. Morgan, Inc. CFDs	Ticker-ID AND/OR Company Name	Yes	Yes
27	Coca-Cola, Co. CFDs	Ticker-ID AND/OR Company Name	Yes	
29	McDonald's, Corp. CFDs	Ticker-ID AND/OR Company Name	Yes	
34	Oracle, Corp. CFDs	Ticker-ID AND/OR Company Name	Yes	

TABLE 31: LIST OF STRING-FILTERED FINANCIAL-INSTRUMENT/TWITTER-FILTER COMBINATIONS FOR WHICH HOURLY CHANGES IN TWITTER DATA LED SECURITIES' HOURLY RETURNS IN A STATISTICALLY-SIGNIFICANT MANNER

As shown in Table 31, the study finds that hourly changes in Twitter message sentiments are able to lead the hourly returns of twelve assets, whilst hourly changes in Twitter message volumes are able to lead the hourly returns of five assets. Furthermore, the hourly returns of four assets were led by both hourly changes in Tweet message sentiments and volumes. Finally, the study found one firm (Bank of America, Corp.) for which hourly changes in Tweet message volumes showed statistically-significant instances of leading the hourly returns of the firm's securities, whilst hourly changes in Tweet message sentiments did not.

As discussed in Chapter 6.3.2.2 and visualised in Figure 21, the study also shows that hourly changes in Twitter message sentiment carry a greater ability to lead the hourly returns of financial securities than hourly changes in Twitter message volumes. This shows that the richer dataset provided by the analytics of Twitter sentiment is more valuable in leading assets' returns than just the Twitter message volumes.

Therefore, the study's findings support Hypothesis Two, showing that that the analysis of Tweets filtered by instrument identifiers and/or company names can be used to lead the returns of individual market-traded securities. In addition, the study's findings also show that message sentiment on individual companies adds information over what is attainable from message volumes (Chapter 6.3.3) – a conclusion which is fortified by the results of the tests of the robustness of the study's findings (Chapter 7.5). Chapter 8.2.2 discusses why message sentiment outperforms message volumes in leading assets' returns.

8.2 Explanations for the study's findings

8.2.1 Why is the information theory measure of dependency a more effective tool for leading assets' returns than linear regression analysis?

The study employs two measures of dependency: linear regression analysis (Chapter 5.3.1) and information theory analysis (Chapter 5.3.2), finding that the former did not show statistically-significant relationships between Twitter data and the returns of market-traded instruments (Chapter 6.2). In contrast, when evaluated using the information theory analysis-based methodology, social media was found to contain statistically-significant lead-time information about the returns of market-traded instruments (Chapter 6.3). A reason for this difference is likely to be the underlying nature of the time-series explored in the study.

The implementation of linear regression analysis as a measure of dependency between two random variables requires a linear (or transformed-linear) relationship between the two[86,87]. This is because linear regression analysis is a normalised covariance, and therefore can only account for any linear relationships[85] between two variables. In contrast, it is widely recognised that financial time-series are strongly nonlinear[88,89], i.e. parameters within the time-series are not a linear function of time. Therefore it is not possible for the nonlinear correlations between the study's dataset of financial and social media data[9,110] to be fully measured using linear regression analysis.

In contrast, information theory can capture both linear and nonlinear dependencies without model specification[85]. Furthermore, information theory has been strongly defended as a measure of predictability and dependence[92]. This is indeed seen in the study's findings (Chapter 6), and in the robustness of the results (Chapter 7), showing the predictability of market data is possible with entropy-based analysis of dependencies with social media data.

A point of note for future works is the inadequacy of linear regression analysis as a measure of dependency for fully capturing the nuances of social media and financial datasets.

8.2.2 Why does message sentiment outperform message volumes in leading assets' returns?

The study's findings show that Tweet sentiments (i.e. its content) contain more information about the future prices of market-traded assets than message volumes. The additional gains from sentiment, over message volumes are detailed in Chapter 6.3.2.2.

This study does not seek to understand the socio-cultural drivers[9] which link social media and internet data with market movements, which at the time of writing are understood to be complex mechanisms requiring in-depth study[13]. However, the Herbert Simon model of decision-making[116] is used by Pries *et al.* (2013)[13] to understand the psychological mechanisms behind interactions between social media data and market movements. The model, which describes the methods involved in logically selecting a path in a decision process, is also applicable to the present study.

The first step of this decision-making model is data acquisition, of which data quality is of importance. For the profit-seeking actor using social media data for market prediction, the sentiment of Tweets relating to a potential investment would therefore be of greater value than just the existence of a message on the topic. This is because message content is data that is of greater *quality* than just message volumes – it carries more information. Indeed as per the study's findings, the former typically does contain more information about market movements than the latter.

It should however be noted that social media data is only one of a multitude of sources which can impact market movements. The logical profit-seeking actor whose decisions are described by the Herbert Simon model will therefore not depend solely on social media data for market insight during the data acquisition phase.

8.2.3 Why does social media data lead the returns of some assets but does not lead the returns of others?

The study identifies that social media leads the returns of some financial assets, but not all. To understand why social media does not lead the returns of all financial assets, an experiment was conducted to determine if the Tweet volumes on companies as filtered by Ticker-ID AND/OR Company Name can be grouped[a].

The k-means clustering algorithm[117] was used to group message volumes on the financial-instrument/Twitter-Filter combinations relating to the ten companies as filtered by Ticker-ID AND/OR Company Name for which the study shows social media data leading the financial data to the 99% level of statistical significance.

The k-means clustering algorithm was configured to group the assets' mean minutely message volumes into two categories, the output of which is shown in Table 32 below.

Filter ID	Instrument name	Mean minutely	k-means cluster	Brand value
15	Google, Inc. CFDs	184	1	$52,132
1	Apple, Inc. CFDs	126.7	1	$87,304
3	Amazon.com, Inc. CFDs	123.1	1	$36,788
29	McDonald's, Corp. CFDs	46.5	2	$21,642
27	Coca-Cola, Co. CFDs	24.8	2	$34,205
22	Intel, Corp. CFDs	12.9	2	$21,139
34	Oracle, Corp. CFDs	5	2	$16,047
8	Cisco Systems, Inc. CFDs	4	2	$15,468
17	The Home Depot, Inc. CFDs	1.9	2	$23,423
25	J.P. Morgan, Inc. CFDs	1.1	2	$13,775

TABLE 32: CLUSTERING OF MEAN MINUTELY MESSAGE VOLUMES OF THE COMPANIES ADMITTED THE STUDY FROM TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME

This exercise identified that mean minutely Tweet message volumes relating Apple, Inc., Amazon.com, Inc. and Google, Inc. are clustered together (with a centroid of 144.1 messages per minute), and are separated from the remaining seven financial-instrument/Twitter-Filter combinations (with a centroid of 12.3 messages per minute).

[a] This grouping exercise did not consider the Tweet volumes associated with the Ticker-ID-only Twitter filter for Apple, Inc. ('$AAPL'), or the string-unfiltered message volumes relating to the S&P500 Index

These three firms (Apple, Inc., Amazon.com, Inc. and Google, Inc.) also have the highest brand values[111].

This clustering shows that the returns of market-traded securities are more predictable with social media analytics for companies with the very highest brand-values, than for global firms with comparatively-lower brand values. This therefore shows that social media analytics for market prediction is currently only suited to a narrow sub-set of high-brand-worth, high-popularity firms. It is not sufficient for a firm to be large or global for social media data to contain an indication of the future returns of its market-traded securities – it must also be of high worldwide popularity.

8.3 Limitations of the study's methodology and suggestions for further work

The study has demonstrated with statistically-significance that the analysis of social media message volumes and sentiments can be used to lead the returns of market-traded securities. The study's limitations are now presented and suggestions for additional technical work are provided.

8.3.1 Limitations in the data

This study is concerned with the analysis of time-series data. There are numerous limitations associated with the acquisition, and processing of both the social media and the financial data used in this study.

Firstly, as discussed in Chapter 4, one of the key technical drawbacks at the time of conducting the study's experiments was the unavailability of historic Twitter datasets in the public domain. Therefore, a large component of the study centred on the collection of Twitter data, first by managing the creation of SocialSTORM (see Chapter 4.1), and then by the creation and use of the Twitter Collection Framework (see Chapter 4.2) – a proprietary framework built for connecting to Twitter's APIs to facilitate the programmatic filtering and downloading of Tweets for the study's experiments. The development durations of both SocialSTORM and the Twitter Collection Framework inherently placed limits on the length of time which was available for the collection of social media data.

However, as discussed in Chapter 5.2, a chronological limit *had* to be set on the length of time available for data-sample collection in order to minimise the effects of routine quarterly updates[a] of ever-changing macroeconomic trends, whilst still offering a range of intra-day market volatilities. Furthermore, the dataset had to be sufficiently small to minimise the effects of seasonality[b] (as discussed in depth in Chapter 5.6.2.1). Finally, and perhaps most importantly, the data collection period had to be small enough to avoid encapsulating significant alternations to the Twitter platform. This is because it has been shown that dramatic alterations to its core product can influence the consistency of Tweet data, driven by the resultant changing demographics of Twitter's users (see Chapter 5.2). It should be noted that past studies in this space do not stipulate a minimum chronological data-size as it is specific to each study – indeed one past work on the analysis of Tweet message sentiments and volumes considered just a 32-day dataset[49].

Therefore, whilst limits did exist on the length of time which was available for the collection of the social media data in the study, the choice of a 3-month dataset collection period based on the carefully-selected criteria was indeed possible.

Provided that the following effects can be mathematically modelled and mitigated, an extension of this study could be performed on a chronologically-larger dataset – this would inevitably provide further insight into the dependencies between social media data and financial data:

- Ability to mitigate the effect of seasonality on Twitter and financial data;
- Ability to mitigate the effect of quarterly macroeconomic trend updates on financial data;
- Ability to mitigate the effect of changes to Twitter's product.

Secondly, further limitations in the study exist from the perspective of Twitter data density. Due to the nature of the License Agreement between Twitter and its users, most programmatic connections to Twitter's APIs provide access of up to 1% of all messages

[a] Macroeconomic data is typically reported on a quarter-by-quarter basis. With reference to this study, the United States Department of Commerce Bureau of Economic analysis, and the UK's Bank of England report macroeconomic data on a quarterly basis.

[b] Seasonality is the effect in time-series data that is driven by economic cycles influenced by the time of year.

passed through its network. As discussed in Chapter 2.1, before the Tweet-collection process began, contractual access for 10% of all messages passed through its network was secured. Thus, whilst the study's 10% dataset is a fully-random sample of the fuller 100% data feed available from Twitter, the analysis of the full feed of all Tweets could provide further insight into social media's ability to lead financial data.

8.3.2 Limitations of the sentiment classification system

This study is built on the analytics of Twitter sentiment using SentiStrength, a transparent dictionary-based classifier which has been shown to consistently outperform baseline competitors in ranking the colloquial nature of user-generated text from internet platforms[50] (see Chapter 5.1). However, the system is only capable of ranking the sentiment on an arbitrary scale of 'negative' to 'positive'. SentiStrength is strongly based on the work of Pennebaker *et al.*[51], which also covers the ability to rank the sentiment of grammatically correct text on additional scales such as: anxiety, optimism, anger, and sadness in their Linguistic Inquiry and Word Count software (LIWC)[a]. Further work in the field of assessing whether social media data can lead financial data should therefore centre on the expansion of the SentiStrength package to incorporate the aforementioned additional scales offered by Pennebaker's LIWC software. This would provide one with the ability to accurately rank the colloquial and often grammatically-incorrect text found in social media using additional mood dimensions. Thus, it is possible that additional insights into whether the sentiment of social media data can lead the returns of financial securities could be ascertained from the analyses of Tweets using these additional mood scales, provided they are adapted to accurately rank informal social media vernacular.

Furthermore, since SentiStrength is only capable of ranking the sentiments of text in English, this study's approach ignores potentially-valuable non-English data passed through Twitter's network. Substantial scope therefore exists for extending the study's approach to the analysis of non-English social media data, provided that SentiStrength's dictionaries can be adapted to rank sentiments in other languages.

[a] http://www.liwc.net/

8.3.3 Limitations of using company names as Twitter filters

As demonstrated by this study in the case of Apple, Inc. CFDs, filtering Tweets by Ticker-ID rather than Ticker-ID AND/OR Company Name shows a stronger ability for Twitter data to lead financial data (see Table 17). This is therefore evidence that filtering Tweets just by company name dilutes social media's predictive powers. This is because using a Twitter filter which mentions a company's name (e.g., "Amazon") does not necessarily guarantee that filtered-in messages will only contain opinions on that firm. The messages can instead contain mentions of a company's service (e.g., "Check out this great deal on Amazon.com") or can in fact be entirely unrelated (e.g., "The Amazon river is unbelievably long"). Thus, whilst this study does demonstrate instances of where social media sentiment filtered by company name leads financial markets in a statistically-significant manner, it is likely that the potential strength of such relationships is diminished by this study's inability to guarantee that Tweets can be filtered to only allow through direct opinions on a company's future performance when filtering by company name. Substantial scope therefore exists for extending the study to only analysing Tweets on a company which contain direct opinions on that firm's future performance. Whilst this is an inherently complex linguistic exercise, such methodologies could employ principles based on advanced part-of-speech tagging[52] methods to infer if a Tweet contains a direct opinion on a firm's future financial performance, or is merely discussing the firm. Such an exercise could provide stronger indications of social media's ability to lead the financial markets.

8.4 Comparison to recent works in the space of market prediction with internet data analytics

The results of this study are used to complement recent studies which seek to predict or track real-world phenomena with social media data (as discussed in Chapter 3.1). Social data have been used to track, predict and measure: epidemiological variables[31,32]; economic variables such as unemployment levels[33], the demand for automobiles[30] and consumer consumption metrics[34]; the popularities and sales of video games, music tracks and feature films[35]; the happiness of internet users as a proxy for the happiness of nations[36]; and the outcomes of political races[37]. Nowcasting has also been used to quantify real-world phenomena in real-time[38] and ahead of the releases of any government-agency data – an endeavour which has been used to track: the present-moment happiness of nations[39,40]; real-time mortality rates[41] and influenza outbreaks[42]; voting intentions during political races[43]; and live macroeconomic activity[44,45].

Of present keen interest is the analysis of social media data for the prediction of financial markets (as discussed in Chapter 3.3). Recent work in this space has used retroactive search-term and parameter-identification methodologies to structure profit-generating social media-driven investment strategies retroactively, typically by only considering message volumes[13-16]. For example, a recent study by Preis *et al.*[13] demonstrated profit-making trading strategies based on the analysis of volumes of particular search terms from Google Trends. However, these works only considered the analysis of social media message volumes (ignoring sentiment), and furthermore were centred on the identification of the search terms which would result in trading strategies which would generate the highest profits retroactively. Such approaches do not describe the quantity of information contained in social media data sentiment on the returns of market-traded securities ahead of time without bias from structuring profit-seeking trading strategies.

This study therefore answers a much more fundamental precursor question to complement and support the aforementioned studies in this field: can the information contained in social media data even lead financial markets, and if so, when? Without using profit-maximisation as the success criterion, and without portfolio structuring and its associated biases, this study demonstrates that social media data contains information about the future returns of market-traded securities.

By using an information theory analysis approach which can capture the nonlinearities of financial and social media datasets, the study shows the extent to which changes in Twitter message volumes can lead the actual or absolute returns of financial instruments, to mirror works which use comparable data sources such as Google Search Trends[13-15] and Yahoo! search engine data[16]. This therefore demonstrates that Twitter message volumes do indeed show statistically-significant instances of leading the returns of market-traded instruments. However, the analysis is extended further by assessing the richer Twitter message sentiment dataset. This demonstrated that the inclusion of sentiment data allows social media analytics to add information over what is attainable from message volumes in leading assets' returns. Furthermore, the study builds on the findings of past works in this space by also showing that a greater message volume per asset indicates the possibility that social media sentiment is more predictive for individual companies; and that a greater message volume per asset indicates the possibility that social media sentiment is more predictive more in advance for individual companies.

Note that the links between social media data and financial market movements are likely the result of complex socio-cultural behaviours[9,13]. Whilst this study does not seek to understand or explain these socio-cultural bridges, it does present quantitative evidence to underpin them and past studies in this area of research. There is therefore a need for in-depth study of the psychological processes involved for full comprehension of market prediction with the analytics of internet data.

9 CONCLUSION AND CONTRIBUTIONS

This chapter concludes the Thesis by reiterating the study's results, and highlighting its contributions.

This study is concerned with the analysis of Twitter messages ('Tweets') to determine the extent to which hourly changes in their sentiments and volumes can lead the hourly returns of market-traded securities. A 3-month dataset is considered, amassed using Twitter's 10% Gardenhose Feed. These Tweets were collected using forty-four Twitter filters which reference twenty-eight market-traded securities.

Linear regression analysis showed no predictive powers (see Chapter 6.2), explained by the nonlinearity of financial and social media data[26,88,89,110]. Instead, the study uses concepts from information theory (see Chapter 5.5) to show statistically-significant lead-time dependencies between Twitter data and the returns of market-traded securities (see Chapter 6.3 for results). Here, the study measures the mutual information between chronologically-offset versions of hourly changes in the sentiment scores and message volumes of Tweets, and hourly changes in the prices of the securities. An evaluation-metric known as information surplus is proposed. It allows for measurement of the extent to which social media data can lead financial data. In such a manner, the study identifies a total of twelve financial-instrument/Twitter-Filter combinations for which social media sentiment contains lead-time information about financial markets. Ten of these represent individual stocks filtered by Ticker-ID AND/OR Company Name, one represents a stock filtered solely by its Ticker-ID ("$AAPL"), and one represents an index (S&P500 Futures).

By applying the information theory analysis methodology to Tweet volumes (rather than Tweet message sentiments), the study also demonstrates that Tweet sentiments lead securities' returns in a statistically-significant manner more often and to a greater extent than Tweet message volumes (as shown in Figure 21). One case (Bank of America, Corp. CFDs) is identified for which Tweet message volumes led the security's returns in a statistically-significant manner whilst Tweet message sentiments did not. The study therefore demonstrates that social media message sentiments add information over what is attainable from social media message volumes.

The study shows that a greater message volume per asset indicates the possibility that social media can be more predictive, and that a greater message volume per asset indicates the possibility that social media can be more predictive more in advance. Testing the sensitivity of the study's results to parameter variation has shown that such findings are robust for a key set of Financial-instrument/Twitter-Filter combinations which are representative of the broad data characteristics of all the assets explored by this study. Based on these observations, it can be stated that the rest of the Financial-instrument/Twitter-Filter combinations from the study's results (as summarised in Table 17) are likely to also exhibit the robustness characteristics (as summarised in List 4). This is because the six key Financial-instrument/Twitter-Filter combinations selected for the parameter-variation experiments encompass the spectrum of data characteristics of the assets in the study's results, as selected by the criteria set out in List 3.

Furthermore, these tests of robustness have shown that additional insight can be gained from social media sentiment for low message-volume assets if using higher-resolution discretisation windows; and that additional insight can be gained from social media by using the Freedman-Diaconis rule for histogram binning in the calculations of mutual information between Twitter data and financial data under certain conditions (as summarised in List 4).

Using k-means clustering (see Chapter 8.2.3), the study also identifies a small number of assets for which Ticker-ID AND/OR Company Name Twitter filters attract a particularly large mean minutely message volume – such messages reference Apple Inc., Google Inc. and Amazon.com Inc., all of which are companies with the highest global brand values. Therefore, any possible trading strategies based on the sentiment analytics of social media data should place emphasis on these high message-volume companies in order to receive the greatest-density "collective wisdom"[47] on a stock's future performance.

In conclusion, this study shows that social media sentiment in a broad-based system like Twitter is indicative of future market movements only on a narrow range of assets, and that such social media sentiments are more indicative than just message volumes. Whilst this study does not seek to understand the complexities of the psychological and socio-cultural mechanisms behind linking internet and social media data to market movements, it does show that sentiment of social-media messages carries more statistically-significant information about future market performance than just the volumes of the messages themselves. This rich data-source should therefore receive further attention using information theory-based analysis, which identified statistically-significant dependencies between social media and financial market data.

9.1 Contributions

This study contributes to an understanding of the real value of Twitter data as a source of information on the future returns of market-traded securities, as an example of predicting/tracking a real-world phenomenon with social media data.

The study’s contributions are:

- An improved understanding of the real value of this new data-source for use as a variable for leading the markets, ascertained without portfolio-structuring bias and without retroactive profit-maximisation as the success criterion as is the case with recent studies[13-15]. This is achieved by the quantification of the amount of information that Twitter data contains about the returns of market-traded securities ahead of time;
- A statistically-significant validation of whether Twitter data can lead the returns of individual market-traded companies and/or stock indices;
- An in-depth insight into the extent to which the quantitative moods of Tweets can lead the markets over and above what is available from the analytics of social media message volumes. This analysis reveals limitations in what can be expected from social media data in leading securities’ returns ahead of time;
- An insight into the generalisations of the extent to which social media message volume can be an indicator of message sentiment being able to lead the returns financial securities and to what extent;
- An insight into the generalisations of the extent to which message sentiment adds predictive powers to message volume when leading the returns of financial securities with social media data;
- The above are achieved via the creation of a series of data collection and analytics frameworks for connection to, and the evaluation of, Twitter data;

- And finally, the conceptual design, management and construction of SocialSTORM – UCL’s Social Media Analytics Engine. As part of this study, SocialSTORM was brought together from conception to realisation at the start of the study for the purposes of providing UCL with access to social media data for research (see Chapter 4.1). Data from SocialSTORM was used for preliminary experiments for this study.

The following papers have been produced in conjunction with this study:

- I. Zheludev, R. Smith and T. Aste. “When Can Social Media Data Lead Financial Markets?”. Sci. Rep. **4**, 4213 (2013);
- R. Wood, I. Zheludev and P. Treleaven. “Mining Social Data with UCL’s SocialSTORM Platform”. DMIN’12 – the 8th International Conference on Data Mining. CSREA Press. 2012. ISBN: 1-60132-208-9.

10 BIBLIOGRAPHY AND REFERENCES

1. Froehlich, F.E. and Kent, A. The Froehlich/Kent Encyclopedia of Telecommunications: Volume 1 - Access Charges in the U.S.A. to Basics of Digital Communications. ISBN: 0824729005. CRC Press (1990).

2. Internet users (per 100 people). *The World Bank* (2013). Database by World Bank on data from the International Telecommunication Union, World Telecommunication/ICT Development Report and database and World Bank estimates. Available at: http://data.worldbank.org/indicator/IT.NET.USER.P2. Accessed: 30/May/2013.

3. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., and Silvestre, B.S. Social media? Get serious! Understanding the functional building blocks of social media. *Bus. Horizons* **54**, 241-251 (2011). Available at: http://dx.doi.org/10.1016/j.bushor.2011.01.005. Accessed: 06/Dec/2014.

4. Twitter Celebrates Initial Public Offering and First Day of Trading on New York Stock Exchange. Press Release by NYSE Euronext (7/Nov/2013). Available at: http://www1.nyse.com/press/1383824955183.html. Accessed: 06/Dec/2014.

5. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J.N. Understanding the Demographics of Twitter Users. Paper presented at the *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain. Menlo Park, CA, USA: The AAAI Press (July 2011). Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234. Accessed: 14/Dec/2014.

6. de Vries, L., Gensler, S., and Leeflang, P.S.H. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *J. Interact. Mark.* **26**, 83-91 (2012). Available at: http://dx.doi.org/10.1016/j.intmar.2012.01.003. Accessed: 06/Dec/2014.

7. Asur, S., and Huberman, B.A. Predicting the Future with Social Media. Paper presented at the *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Theory*, Toronto, Canada. Washington, DC,

USA: IEEE Computer Society (August 2010). Available at: http://dx.doi.org/10.1109/WI-IAT.2010.63. Accessed: 14/Dec/2014.

8. O'Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. From tweets to polls: Linking text *sentiment* to public opinion time series. Paper presented at the *Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC, USA. Menlo Park, CA, USA: The AAAI Press (May 2010). Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842. Accessed: 14/Dec/2014.

9. Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *J. Comp. Sci.* **2**, 1-8 (2011). Available at: http://dx.doi.org/10.1016/j.jocs.2010.12.007. Accessed: 06/Dec/2014.

10. Zhang, X., Fuehres, H., and Gloor, P.A. Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia Soc. Behav. Sci.* **26**, 55-62 (2011). Available at: http://dx.doi.org/10.1016/j.sbspro.2011.10.562. Accessed: 06/Dec/2014.

11. Mao, Y., Wei, W., Wang, B., and Liu, B. Correlating S&P500 stocks with Twitter data. Paper presented at the *First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, Beijing, China. New York, NY, USA: ACM (August 2012). Available at: http://dx.doi.org/10.1145/2392622.2392634. Accessed: 14/Dec/2014.

12. Ruiz, E., J, Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. Correlating Financial Time Series with Micro Blogging Activity. Paper presented at the *Fifth ACM International Conference on Web search and Data Mining*, Seattle, WA, USA. New York, NY, USA: ACM (February 2012). Available at: http://dx.doi.org/10.1145/2124295.2124358. Accessed: 14/Dec/2014.

13. Preis, T., Moat, H.S., and Stanley, H.E. Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci. Rep.* **3**, 1684 (2013). Available at: http://dx.doi.org/10.1038/srep01684. Accessed: 06/Dec/2014.

14. Challet, D., and Bel Hadj Ayed, A. Predicting financial markets with Google Trends and not so random keywords. *arXiv preprint arXiv:1307.4643v3* (2013). Available at: http://arxiv.org/abs/1307.4643v3. Accessed: 06/Dec/2014.

15. Preis, T., Reith, D., and Stanley, H.E. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philos. T. R. Soc. A.* **368**, 5707-5719 (2010). Available at: http://dx.doi.org/10.1098/rsta.2010.0284. Accessed: 06/Dec/2014.

16. Bordino, I., Battiston, S., Caldarelli, G., and Cristelli, M. Web search queries can predict stock market volumes. *PloS one* **7**, e40014 (2012). Available at: http://dx.doi.org/10.1371/journal.pone.0040014. Accessed: 06/Dec/2014

17. Moat, H.S., Curme, C., Avakian, A, Kenett, D.Y., Stanley, H.E., and Preis, T. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Sci. Rep.* **3**, 1801 (2013). Available at: http://dx.doi.org/10.1038/srep01801. Accessed: 06/Dec/2014.

18. Grossman, S.J., and Stiglitz, J.E. On the Impossibility of Informationally Efficient Markets. *Am. Econ. Rev.* **70**, 393-408 (1980). Available at: http://www.jstor.org/stable/1805228. Accessed: 06/Dec/2014.

19. Brody, D., Meister, B., and Parry, M. Informational inefficiency in financial markets. *Math. Fin. Econ.* **6**, 249-259 (2012). Available at: http://dx.doi.org/10.1007/s11579-012-0078-1. Accessed: 06/Dec/2014.

20. Mackintosh, J. Last tweet for Derwent's Absolute Return. *The Financial Times* (24/May/2012). News article. Available at: http://www.ft.com/cms/s/0/d5d9c3f8-a5bf-11e1-b77a-00144feabdc0.html. Accessed: 07/Aug/2014.

21. Mackintosh, J. Traders tap Twitter for top stock tips. *The Financial Times* (11/Oct/2013). News article. Available at: http://www.ft.com/cms/s/0/e1335c66-3284-11e3-91d2-00144feab7de.html. Accessed: 07/Aug/2014.

22. Alden, W. Separating the Market-Moving Tweets From the Chaff. *The New York Times* (11/Nov/2013). News article. Available at: http://dealbook.nytimes.com/2013/11/11/separating-the-market-moving-tweets-from-the-chaff/?_php=true&_type=blogs&_r=0. Accessed: 07/Aug/2014.

23. Wells, G. Tools Help Investors Wade Through All the Chatter on Twitter. *The Wall Street Journal* (11/Dec/2013). Available at: http://online.wsj.com/news/articles/SB10001424052702303653004579210260999153206. Accessed: 07/Aug/2014.

24. Zheludev, I. Seeing the future with social media. Talk presented at *TEDxUCL*, London, UK (02/Jun/2012). Presentation. Available at: http://tedxtalks.ted.com/video/TEDxUCL-Ilya-Zheludev-Seeing-th;search%3Atag%3A%22tedxucl%22. Accessed: 14/Dec/2013.

25. Naaman, M., Boase, J., and Lai, C. Is it all About Me?: message content in social awareness streams. Paper presented at the *2010 ACM conference on Computer supported cooperative work*, Savannah, GA, USA. New York, NY, USA: ACM (February 2010). Available at: http://dx.doi.org/10.1145/1718918.1718953. Accessed: 14/Dec/2014.

26. Bollen, J., Pepe, Al, and Mao, H. Modelling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Paper presented at the *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain. Menlo Park, CA, USA: The AAAI Press (July 2011). Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2826/3237. Accessed: 14/Dec/2014.

27. Manning, C.D., and Schütze, H. Foundations of statistical natural language processing. ISBN: 0262133601. MIT Press (July 1999).

28. Turney, P.D. Thumbs Up or Thumbs Down?: Semantic orientation applied to unsupervised classification of reviews. Paper presented at the *Fortieth annual meeting on association for computational linguistics*. Philadelphia, PA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics (July 2002). Available at: http://dx.doi.org/10.3115/1073083.1073153. Accessed: 14/Dec/2014.

29. Kleinnijenhuis, J., Schultz, F., Oegema, D., and Atteveldt, W.V. Financial news and market panics in the age of high-frequency sentiment trading algorithms. *Journalism* **14**, 271-291 (2013). Available at: http://dx.doi.org/10.1177/1464884912468375. Accessed: 06/Dec/2014.

30. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., and Danforth, C.M. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE* **6**, e26752 (2011). Available at: http://dx.doi.org/10.1371/journal.pone.0026752. Accessed: 06/Dec/2014.

31. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D., and Weinstein, R.A. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **47**, 1443-1448 (2008). Available at: http://dx.doi.org/10.1086/593098. Accessed: 06/Dec/2014.

32. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012-1014 (2009). Available at: http://dx.doi.org/10.1038/nature07634. Accessed: 06/Dec/2014.

33. Baker, S., and Fradkin, A. What drives job search? Evidence from Google search data. *Stanford Institute for Economic Policy Research*, Stanford, CT, USA (2011). Stanford Institute for Economic Policy Research Discussion Paper No. 10-020. Available at: http://www-siepr.stanford.edu/RePEc/sip/10-020.pdf. Accessed: 12/May/2014.

34. Vosten, S., and Schmidt, T. Forecasting private consumption: survey-based indicators vs. Google trends. *J. Forecasting* **30**, 565-578 (2011). Available at: http://dx.doi.org/10.1002/for.1213. Accessed: 14/Dec/2014.

35. Goel, S., Hofman, J.M, Lahaie, S., Pennock, D.M, and Watts, D.J. *Proc. Natl. Acad. Sci. U.S.A* **107**, 17486-17490 (2010). Available at: http://dx.doi.org/10.1002/for.1213. Accessed: 06/Dec/2014.

36. Brill, E. A simple rule-based part of speech tagger. Paper presented at the *Third Conference on Applied Natural Language Processing*, Trento, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics (March 1992). Available at: http://dx.doi.org/10.3115/1075527.1075553. Accessed: 14/Dec/2014.

37. Choi, H., and Varian, H. Predicting the Present with Google Trends. *Econ. Rec.* **88**, 2-9 (2012). Available at: http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x. Accessed: 06/Dec/2014.

38. Lampos, V., and Cristianini, N. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.* **3**, 72 (2012). Available at: http://dx.doi.org/10.1145/2337542.2337557. Accessed: 06/Dec/2014.

39. Lansdall-Welfare, T., Lampos, V., and Cristianini, N. Nowcasting the mood of the nation. *Significance* **9**, 26-28 (2012). Available at: http://dx.doi.org/10.1111/j.1740-9713.2012.00588.x. Accessed: 06/Dec/2014.

40. Lampos, V., Lansdall-Welfare, T., Araya, R., and Cristianini, N. Analysing mood patterns in the united kingdom through twitter content. *arXiv preprint arXiv:1304.5507* (2013). Available at: http://arxiv.org/abs/1304.5507. Accessed: 06/Dec/2014.

41. Green, H.K., Andrews, N.J., Bickler, G., and Pebody, R.G. Rapid estimation of excess mortality: nowcasting during the heatwave alert in England and Wales in June 2011. *J. Epidemiol. Commun. H.* **66**, 866-868 (2012). Available at: http://dx.doi.org/10.1136/jech-2011-200962. Accessed: 06/Dec/2014.

42. Ray, J., and Brownstein, J.S. Nowcasting influenza outbreaks using open-source media reports. *Sandia National Laboratories*, Livermore, CA, USA (2013). Sandia Technical Report, SAND2013-0963. Available at: http://www.sandia.gov/~jairay/Presentations/sand2013-0963.pdf. Accessed: 12/May/2014.

43. Lampos, V. On voting intentions from Twitter content: a case study on UK 2010 General Election. *arXiv preprint arXiv:1204.0423* (2012). Available at: http://arxiv.org/abs/1204.0423. Accessed: 06/Dec/2014.

44. Aruoba, S.B., and Diebold, F.X. Real-time macroeconomic monitoring: Real activity, inflation and interactions. *National Bureau of Economic Research*, Cambridge, MA, USA (2010). National Bureau of Economic Research Working Paper No. 15657. Available at: http://www.nber.org/papers/w15657. Accessed: 12/May/2014.

45. Kholodilin, K.A., Podstawski, M., and Siliverstovs, B. Do Google searches help in nowcasting private consumption? A real-time evidence for the US. *KOF Swiss Economic Institute*, Zurich, Switzerland (2010). German Institute for Economic Research Working Paper No. 256. Available at:

http://www.kof.ethz.ch/en/publications/p/kof-working-papers/256/. Accessed: 12/May/2014.

46. Facebook Demographics and Statistics Report 2010. *iStrategyLabs* (2010). Dataset on the demographics of Facebook users. Available at: http://istrategylabs.com/2010/01/facebook-demographics-and-statistics-report-2010-145-growth-in-1-year/. Accessed: 11/Dec/2012.

47. Saavedra, S., Duch, J., and Uzzi, B. Tracking Traders' Understanding of the Market Using e-Communication Data. *PLoS ONE* **6**, e26705 (2011). Available at: http://dx.doi.org/10.1371/journal.pone.0026705. Accessed: 06/Dec/2014.

48. Galton, F. Vox Populi. *Nature*. **75**, 450-451 (1907). Available at: http://dx.doi.org/10.1038/075450a0. Accessed: 21/Dec/2014.

49. Oliveira, N., Cortez, P., and Areal, N. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. Paper presented at the *Third International Conference on Web Intelligence, Mining and Semantics*, Madrid, Spain. New York, NY, USA: ACM (June 2013). Available at: http://dx.doi.org/10.1145/2479787.2479811. Accessed: 14/Dec/2014.

50. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Imf. Sci. Technol.* **61**, 2544-2558 (2010). Available at: http://dx.doi.org/10.1002/asi.21416. Accessed: 06/Dec/2014.

51. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., and Booth, R.J. The development and psychometric properties of LIWC2007. *The University of Texas at Austin* (2007). Instruction Manual. Available at: http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_LanguageManual.pdf. Accessed: 15/Jan/2013.

52. Mitkov, R. The Oxford Handbook of Computational Linguistics. ISBN: 019927634X. Oxford University Press (March 2005).

53. Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. Predicting elections with twitter: What 140 characters reveal about political sentiment. Paper presented at the *Fourth International AAAI Conference on Weblogs and*

Social Media, Washington, DC, USA. Menlo Park, CA, USA: The AAAI Press (May 2010). Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852. Accessed: 14/Dec/2014.

54. Diakopoulos, N.A., and Shamma, D.A. Characterizing debate performance via aggregated twitter sentiment. Paper presented at *CHI '10 SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, GA, USA. New York, NY, USA: ACM (April 2010). Available at: http://dx.doi.org/10.1145/1753326.1753504. Accessed: 14/Dec/2014.

55. Nofsinger, J.R. Social Mood and Financial Economics. *J. Behav. Finance* **6**, 144-160 (2005). Available at: http://dx.doi.org/10.1207/s15427579jpfm0603_4. Accessed: 06/Dec/2014.

56. Sukumar, N. DCM Capital Puts Itself Up for Sale in Online Auction. *Bloomberg* (04/Feb/2013). News article. Available at: http://www.bloomberg.com/news/2013-02-04/dcm-capital-puts-itself-up-for-sale-in-online-auction.html. Accessed: 06/Dec/2014.

57. Wood, R., Zheludev, I., and Treleaven, P. Mining Social Data with UCL's Social Media Platform. Paper presented at the *2012 International Conference on Data Mining*, Las Vegas, NV, USA. Las Vegas, NV, USA: CSREA Press (July 2012). Available at: http://worldcomp-proceedings.com/proc/p2012/DMI9011.pdf. Accessed: 14/Dec/2014.

58. Knight, S. School for Quants. *The Financial Times* (02/Mar/2012). News article. Available at: http://www.ft.com/cms/s/2/0664cd92-6277-11e1-872e-00144feabdc0.html. Accessed: 07/Aug/2014.

59. Costolo, R. Registration Statement for Twitter, Inc. *Securities and Exchange Commission* (03/Oct/2013). Form S-1 Registration Statement under the Securities Act of 1933. Available at: http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm. Accessed: 08/Dec/2014.

60. Batool, R., Khattak, A.M., Maqbool J., and Sungyoung, L. Precise tweet classification and sentiment analysis. Paper presented at the *2013 IEEE/ACIS*

12th International Conference on Computer and Information Science (ICIS), Niigata, Japan: IEEE (June 2013). Available at: http://dx.doi.org/10.1109/ICIS.2013.6607883. Accessed: 14/Dec/2014.

61. Kouloumpis, E., Wilson, T., and Moore, J. Twitter sentiment analysis: The good the bad and the omg!. Paper presented at the *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain. Menlo Park, CA, USA: The AAAI Press (July 2011). Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2857/3251. Accessed: 14/Dec/2014.

62. Long, M. Sentiment Analysis using a Naive Bayes classifier. Thesis at *University College London* for the MSc in Computer Science (2012).

63. Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment in Twitter events. *J. Am. Soc. Imf. Sci. Technol*. **62**, 406-418 (2011). Available at: http://dx.doi.org/10.1002/asi.21462. Accessed: 06/Dec/2014.

64. Heitner, D. Verizon's Super Bowl Scheme Is To Save £4 Million And Light Up The Sky. *Forbes* (30/Jan/2014). News article. Available at: http://www.forbes.com/sites/darrenheitner/2014/01/30/verizons-super-bowl-scheme-is-to-save-4-million-and-light-up-the-sky/. Accessed: 14/May/2014.

65. Grossman, S. Want to Light Up the London Eye? Just Tweet That the Olympics are 'Totes Amazeballs'. *Time Magazine* (27/Jul/2012). News article. Available at: http://olympics.time.com/2012/07/27/want-to-light-up-the-london-eye-just-tweet-that-the-olympics-are-totes-amazeballs/?xid=rss-topstories. Accessed: 12/May/2014.

66. Witten, I.H., and Frank, E. Data Mining: Practical machine learning tools and techniques. ISBN: 008047702X. Morgan Kaufmann Publishers (July 2005).

67. Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Imf. Sci. Technol.* **63**, 163-173 (2012). Available at: http://dx.doi.org/10.1002/asi.21662. Accessed: 06/Dec/2014.

68. Engle, R.F., Ghysels, E., and Sohn, B. Stock market volatility and macroeconomic fundamentals. *Rev. Econ. Stat.* **95**, 776-797 (2013). Available at: http://dx.doi.org/10.1162/REST_a_00300. Accessed: 06/Dec/2014.

69. Cakmakli, C., and van Dijk, D. Getting the most out of macroeconomic information for predicting stock returns and volatility. *Erasmus University Rotterdam* (22/Nov/2010). Tinbergen Institute discussion Paper 2010-115/4. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1713687. Accessed: 12/May/2014.

70. Bellona, D. Coming soon: a whole new you, in your Twitter profile. *Twitter* (8/Apr/2014). Blog post. Available at: https://blog.twitter.com/2014/coming-soon-a-whole-new-you-in-your-twitter-profile. Accessed: 12/May/2014.

71. Williams, R. Twitter rolls out new profile design to all users. *The Telegraph* (23/Apr/2014). News article. Available at: http://www.telegraph.co.uk/technology/twitter/10782237/Twitter-rolls-out-new-profile-design-to-all-users.html. Accessed: 12/May/2014.

72. Bank of England. Quarterly Bulletin Economic Statistics Articles. *Bank of England*. Data on economic performance. Available at: http://www.bankofengland.co.uk/publications/Pages/quarterlybulletin/econstat.aspx. Accessed: 12/May/2014.

73. U.S. Department of Commerce, Bureau of Economic Analysis. National Economic Accounts. *U.S. Department of Commerce*. Data on economic performance. Available at: http://www.bea.gov/national/index.htm. Accessed: 12/May/2014.

74. Results from The Twitter Blog for: December 2012. *Twitter* (Dec/2012). Blog post. Available at: https://blog.twitter.com/?year=2012&month=12. Accessed: 12/May/2014.

75. Results from The Twitter Blog for: January 2013. *Twitter* (Jan/2013). Blog post. Available at: https://blog.twitter.com/?year=2013&month=1. Accessed: 12/May/2014.

76. Results from The Twitter Blog for: February 2013. *Twitter* (Feb/2013). Blog post. Available at: https://blog.twitter.com/?year=2013&month=2. Accessed: 12/May/2014.

77. Results from The Twitter Blog for: March 2013. *Twitter* (Mar/2013). Blog post. Available at: https://blog.twitter.com/?year=2013&month=3. Accessed: 12/May/2014.

78. Chatfield, C. Time-Series Forecasting. ISBN: 978-1584880639. Chapman and Hall/CRC (October 2000).

79. Andersen, T.G., Bollerslev, T., and Lange, S. Forecasting financial market volatility: Sample frequency vis-à-vis forecast horizon. *J. Empir. Financ.* **6**, 457-477 (1999). Available at: http://dx.doi.org/10.1016/S0927-5398(99)00013-4. Accessed: 06/Dec/2014.

80. Zheludev, I. First Year Viva Report: Predicting Financial Markets and Consumer Confidence with Social Media. First Year Viva Report at *University College London* for the PhD in Financial Computing (2012).

81. Zheludev, I. Research with SocialSTORM: Overview on SocialSTORM. Presentation given to *Fulcrum Asset Management*, London, UK (March 2012).

82. Zheludev, I. MPhil to PhD Transfer Report: Predicting Financial Markets with Social Media. MPhil to PhD Transfer Report at *University College London* for the PhD in Financial Computing (2013).

83. Stigler, S.M. Francis Galton's Account of the Invention of Correlation. *Statist. Sci.* **4**, 73-79 (1989). Available at: http://www.jstor.org/stable/2245329. Accessed: 06/Dec/2014.

84. Rodgers, J.L., and Nicewander, A. Thirteen Ways to Look at the Correlation Coefficient. *Am. Stat.* **42**, 59-66 (1988). Available at: http://dx.doi.org/10.1080/00031305.1988.10475524. Accessed: 06/Dec/2013.

85. Dionisio, A., Menezes, R., and Mendes, D.A. Mutual information: a measure of dependency for nonlinear time series. *Phys. A.* **344**, 326-329 (2004). Available at: http://dx.doi.org/10.1016/j.physa.2004.06.144. Accessed: 06/Dec/2014.

86. Granger, C. and Lin, J-L. Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Ser. Anal.***15**, 371-384 (1994). Available at: http://dx.doi.org/10.1111/j.1467-9892.1994.tb00200.x. Accessed: 13/Dec/14.

87. Bernhard, H-P., and Darbellay, G.A. Performance analysis of the mutual information function for nonlinear and linear signal processing. Paper presented

at the *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, USA: IEEE (March 1999). Available at: http://dx.doi.org/10.1109/ICASSP.1999.756217. Accessed: 13/Dec/2014.

88. Cao, L.J., and Tay, F.E.H. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Networks*. **14**, 1506-1518 (2003). Available at: http://dx.doi.org/10.1109/TNN.2003.820556. Accessed: 13/Dec/2014.

89. Qi, M. and Maddala, G.S. Economic factors and the stock market: a new perspective. *J. Forecasting*. **18**, 151-166. Available at: http://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291099-131X%28199905%2918:3%3C151::AID-FOR716%3E3.0.CO;2-V/abstract. Accessed: 13/Dec/2014.

90. Anscombe, F.J. Graphs in Statistical Analysis. *Am. Stat.* **27**, 17-21 (1973). Available at: http://dx.doi.org/10.1080/00031305.1973.10478966. Accessed: 06/Dec/2014.

91. Reza, F. An Introduction to Information Theory (Dover Books on Mathematics). ISBN: 0486682102. Dover Publications (June 2010).

92. Urbach, R.M.A. Footprints of Chaos in the Markets: Analyzing non-linear time series in financiual markets and other real systems. ISBN: 0273635735. Financial Times / Prentice Hall (December 1999).

93. Ash, R. Information Theory (Dover Books on Mathematics). ISBN: 0486665216. Dover Publications (November 1990).

94. Scott, D.W. Multivariate Density Estimation: Theory, Practice and Visualization. ISBN: 9780470317686. Wiley (September 2009).

95. Sturges, H.A. The Choice of a Class Interval. *JASA* **21**, 65-66 (1926). Available at: http://www.jstor.org/stable/2965501. Accessed: 06/Dec/2014.

96. Wand, M.P. Data-based choice of histogram bin width. *Am. Stat*. **51**, 59-64 (1997). Available at: http://dx.doi.org/10.1080/00031305.1997.10473591. Accessed: 06/Dec/2014.

97. Han, J., and Kamber, M. Data Mining: Concepts and Techniques. ISBN: 9780123814807. Elsevier (June 2011).

98. Madsen, H. Time Series Analysis. ISBN: 9781420059670. CRC Press (November 2007).

99. Kitagawa, G. Introduction to Time Series Modeling. ISBN: 9781584889229. CRC Press (June 2010).

100. Diebold, F. Elements of Forecasting. ISBN: 9781285414416. Cengage Learning (December 2006).

101. Hyndman, R.J. Forecasting: principles and practice. ISBN: 0987507109. OTexts (October 2013).

102. Miller, K.S., and Rochwarger, M. A covariance approach to spectral moment estimation. *IEEE Trans. Inf. Theory*. **18**, 588-596. Available at: http://dx.doi.org/10.1109/TIT.1972.1054886. Accessed: 06/Dec/2014.

103. Zwillinger, D. CRC Standard Mathematical Tables and Formulae, 31st Edition (Discrete Mathematics and Its Applications). ISBN: 1584882913. Chapman and Hall/CRC (November 2002).

104. Hamilton, J.D. Time series analysis. ISBN: 0691042896. Princeton University Press (January 1994).

105. Dunn, P.F. Measurement and Data Analysis for Engineering and Science, Second Edition. ISBN: 9781439825693. CRC Press (January 2010).

106. Boashash, B. Time-frequency signal analysis. ISBN: 0130074446. Elsevier (October 2003).

107. Wei, W.W. Time series analysis: Univariate and Multivariate Methods. ISBN: 0321322169. Pearson Addison Wesley (July 2005).

108. Christmann, E.P., and Badgett, J.L. Interpreting Assessment Data: Statistical Techniques You Can Use. ISBN: 9781933531366. NSTA Press (October 2008).

109. Dancey, C.P., and Reidy, J. Statistics Without Maths for Psychology: Using SPSS for Windows. ISBN: 9780131249417. Prentice Hall (April 2011).

110. Mittal, A, and Goel, A. Stock prediction using twitter sentiment analysis. *Stanford University*, CA, USA (2012). Stanford University Working Paper. Available at: ftp://cse.shirazu.ac.ir:5001/Tempbuffer/InGruheKhashen/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf. Accessed: 13/Dec/2014.

111. D’Souza, S. Brandirectory Global 500 2013 Top Brands. *Brandirectory* (2013). Database on global brand popularity. Available at: http://brandirectory.com/league_tables/table/global-500-2013. Accessed: 29/Oct/2013.

112. Aldridge, I. FINalternatives Survey: High-Frequency Trading Has a Bridge Future. *FINalternatives* (22/Jul/2009). News article. Available at: http://www.finalternatives.com/node/8583. Accessed: 16/Jul/2014.

113. Birge, L,. and Rozenholc, Y. How many bins should be put in a regular histogram. *ESAIM, Probab. Stat.* **10**, 24-45 (2006). Available at: http://dx.doi.org/10.1051/ps:2006001. Accessed: 06/Dec/2014.

114. Tukey, J.W. Exploratory data analysis. ISBN: 0201076160. Pearson (January 1977).

115. Freedman, D., and Diaconis, P. On the histogram as a density estimator:L_2 theory. *Probab. Theory Related Fields* **57**, 453-476 (1981). Available at: http://dx.doi.org/10.1007/BF01025868. Accessed: 06/Dec/2013.

116. Simon, H. A. A behavioral model of rational choice. *Q. J. Econ.* **69**, 99-118 (1955). Available at: http://www.jstor.org/stable/1884852. Accessed: 14/Dec/2014.

117. MacQueen, J.B. Some Methods for Classification and Analysis of MultiVariate Observations. Paper presented at the *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA. Berkeley, CA, USA: University of California Press (June 1965). Available at: http://www-m9.ma.tum.de/foswiki/pub/WS2010/CombOptSem/kMeans.pdf. Accessed: 14/Dec/2014.

11 APPENDIX

11.1 Financial-instrument/Twitter-Filter combinations for which social media leads financial data to the 99% level of statistical significance

11.1.1 Apple, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Apple, Inc. is an electronic equipment manufacturer headquartered in California, USA and is listed in the NASDAQ 100 stock index with a market capitalisation of $641bn as at December 2014[a], and is the world's highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the design and production of consumer-orientated mobile computing and mobile telephony hardware and software.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$AAPL" AND/OR "Apple", to capture Tweets mentioning Apple, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 16.8 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Apple, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated two time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. Both time-shifts were attributed to hourly changes in Twitter's negative sentiments on the company.

As seen in Figure 27 below, hourly changes in neither the net sentiments nor the positive sentiments showed any ability to lead Apple, Inc. CFD's hourly returns in a statistically-significant manner.

[a] http://uk.finance.yahoo.com/q/pr?s=AAPL

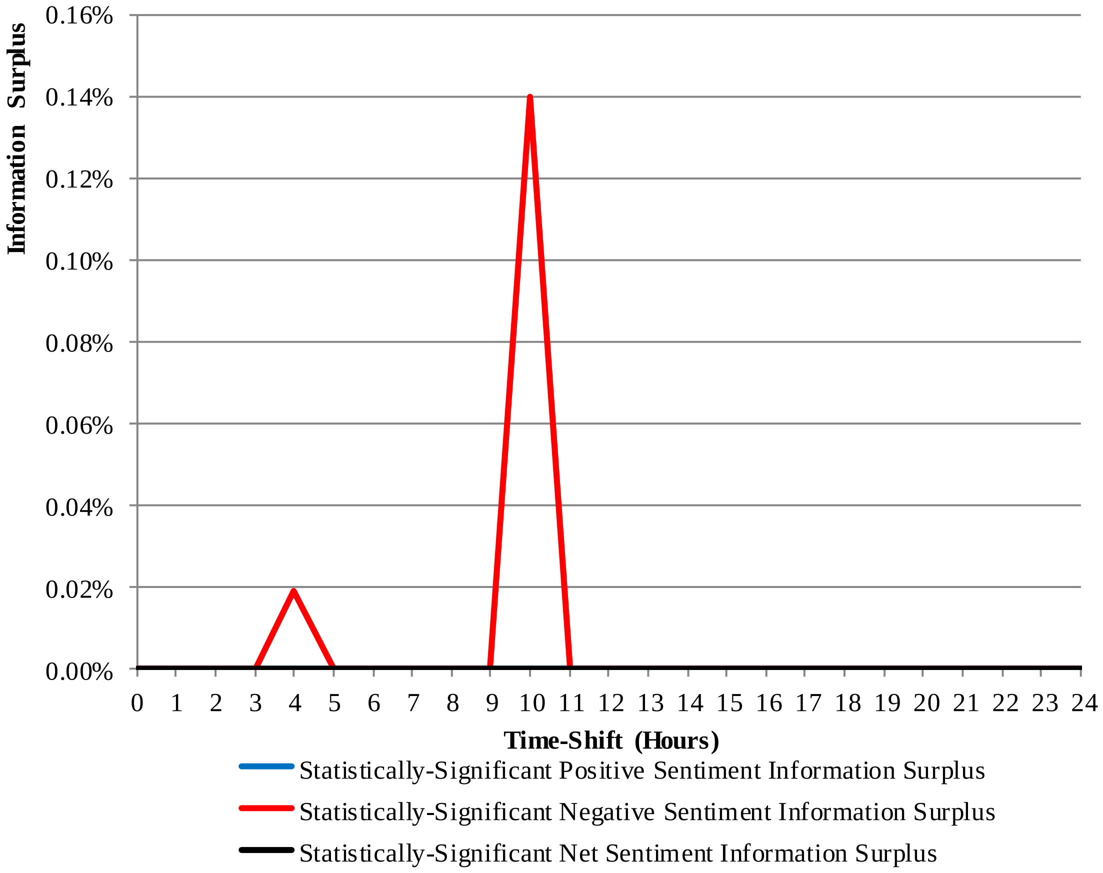


FIGURE 27: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF APPLE, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 27 shows that hourly changes in Twitter sentiment demonstrated its greatest ability to lead the asset's hourly returns at a leading time-shift of 10-hours, with a peak information surplus of 0.14%.

Hourly changes in Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute returns of Apple, Inc. CFDs in a statistically-significant manner.

11.1.2 Apple, Inc. CFDs, with social media data filtered by Ticker-ID only

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$AAPL", to capture Tweets mentioning Apple, Inc.'s Ticker-ID only. The company in question is the same as seen in Chapter 11.1.1. In this manner, 237 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Apple, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated two time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. Similarly to what was observed with the Apple Ticker-ID AND/OR Company Name Twitter-Filter (Chapter 11.1.1), both time-shifts were attributed to hourly changes in Twitter's negative sentiments on the company. As seen in Figure 28 below, hourly changes in neither the net sentiments nor the positive sentiments showed any ability to lead Apple, Inc. CFD's hourly returns in a statistically-significant manner with the Ticker-ID only Twitter-Filter.

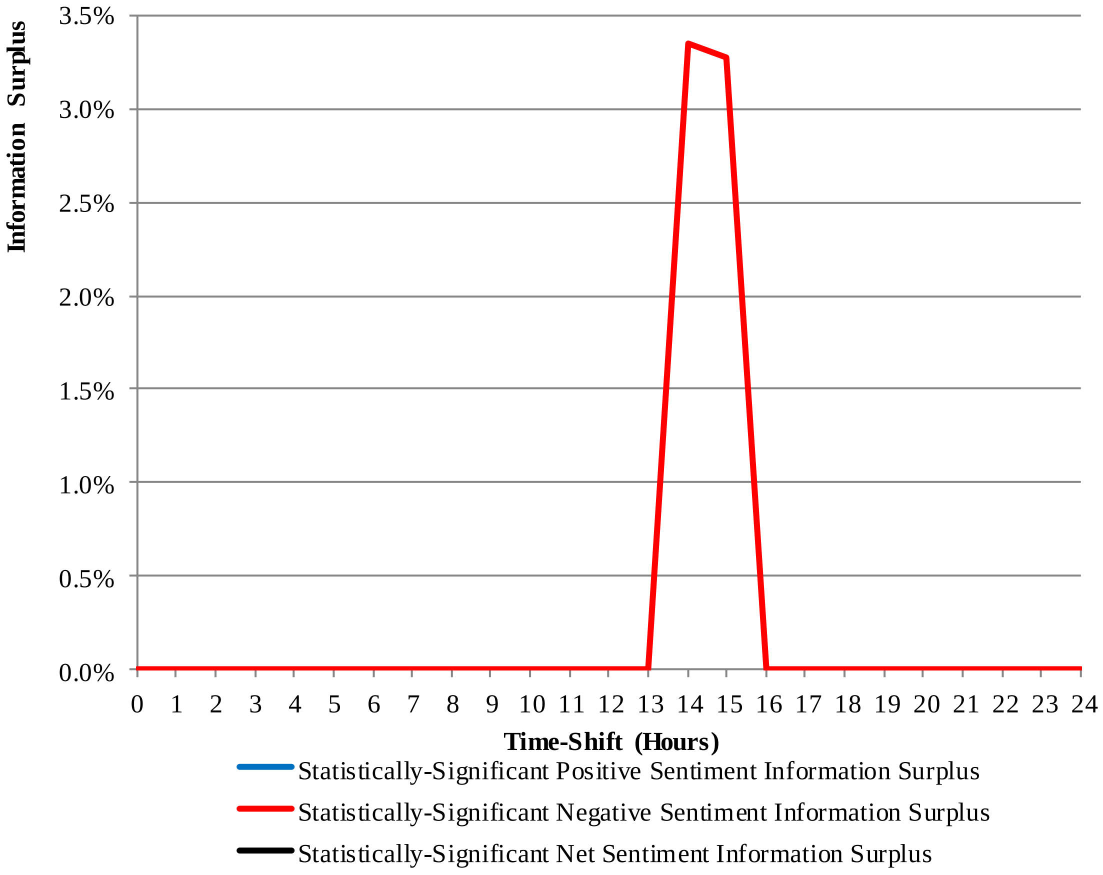


FIGURE 28: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID-ONLY LED THE HOURLY RETURNS OF APPLE, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 28 shows that hourly changes in Twitter's negative sentiment data demonstrated the greatest ability to lead the asset's hourly returns in advance at a leading time-shift of 14-hours, with a peak information surplus of 3.35

In addition, this financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Tweet message volumes led the asset's hourly returns ahead of time in a statistically-significant manner. As seen in Figure 29 below, hourly changes in Tweet message volumes led the hourly returns of Apple, Inc. CFDs for five time-shifts, with a peak information surplus of 0.89% occurring at a time-shift of 2-hours. Hourly changes in Tweet message volumes led the absolute returns of Apple, Inc. CFDs for nine time-shifts, with a peak information surplus of 0.94% occurring at a leading time-shift of 2-hours. This indicates that hourly changes in Tweet message volumes show a greater capacity to lead the asset's absolute hourly returns than the asset's actual hourly returns. However, hourly changes in Twitter sentiment generate a peak information surplus that is greater than the peak information surplus generated from hourly changes in Twitter message volumes. This indicates that hourly changes in social media sentiment carry a greater ability to lead this asset's hourly returns than hourly changes in social media message volumes.

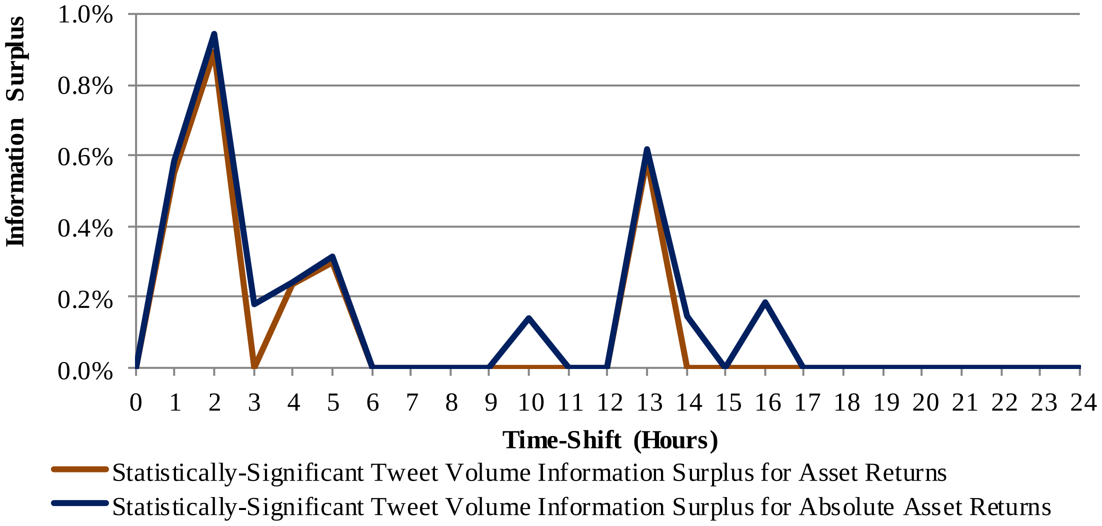


Figure 29: Time-shifts at which hourly changes in Tweet message volumes filtered by Ticker-ID-only led the hourly returns of Apple, Inc. CFDs in a statistically-significant manner

11.1.3 Amazon.com, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Amazon.com, Inc. is an internet-based catalogue and mail-order business headquartered in Washington, USA and is listed in the NASDAQ 100 stock index and the AMEX internet index with a market capitalisation of $138bn as at December 2014[a], and is the world's 10th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the operation of online-retail websites and product shipping, cloud-based internet hosting, and online digital distribution of professionally-published text, video and audio media.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$AMZN" AND/OR "Amazon" (one of Amazon.com, Inc.'s trading names), to capture Tweets mentioning Amazon.com, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 16.2 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Amazon.com, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 30 below, hourly changes in Twitter's negative sentiments on the company were able to lead the returns of Amazon.com, Inc. CFDs for twenty time-shifts, with a peak information surplus of 1.91% occurring at a leading time-shift of 12-hours. Hourly changes in Twitter's positive sentiments on the company were able to lead the hourly returns of its CFDs for one time-shift, with a peak information surplus of 2.59% at a leading time-shift of 19-hours. Finally, hourly changes in Twitter's net sentiments on the company were able to lead the hourly returns of its CFDs for nine time-shifts, with a peak information surplus of 3.47% at a leading time-shift of 20-hours, indicating that hourly changes in Twitter's net sentiment on Amazon.com are most indicative of the hourly returns of the asset's CFDs ahead of time.

[a] http://uk.finance.yahoo.com/q/pr?s=AMZN

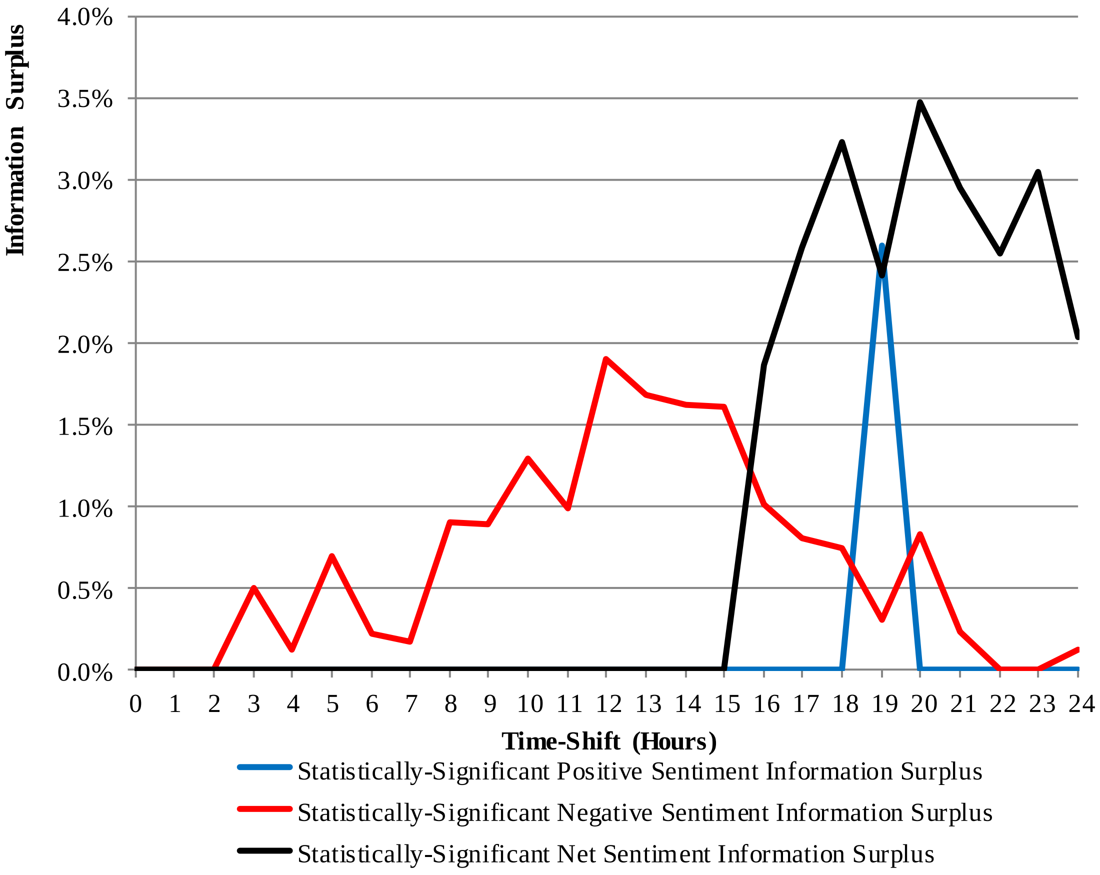


FIGURE 30: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF AMAZON.COM, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 30 shows the range of statistically-significant information surplus values for the three sentiment types.

Hourly changes in Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of Amazon.com, Inc. CFDs in a statistically-significant manner.

11.1.4 Changes in Twitter message volumes lead the returns of Bank of America, Corp. CFDs, but changes in Tweet message sentiments do not when social media data are filtered by Ticker-ID AND/OR Company Name

Bank of America, Corp. is a financial services provider, as detailed in Chapter 6.3.1.1.

This financial-instrument/Twitter-Filter combination demonstrated no instances of when hourly changes in the Tweet message sentiments (whether positive, negative or net) was able to lead the asset's hourly returns ahead of time in a statistically-significant manner. However, hourly changes in the Tweet message volumes did show the ability to lead the hourly returns and the absolute hourly returns of Bank of America, Corp. CFDs in a statistically-significant manner.

As seen in Figure 31 below, hourly changes in the Tweet message volumes led the hourly returns of Bank of America, Corp. CFDs for eight time-shifts, with a peak information surplus of 0.61% occurring at a time-shift of 1-hour. Hourly changes in the Tweet message volumes led the absolute hourly returns of Bank of America, Corp. CFDs also for eight time-shifts, with a peak information surplus of 0.65% occurring at a leading time-shift of 14-hours. This indicates that hourly changes in Tweet message volumes show a greater capacity to lead the asset's absolute hourly returns than the asset's actual hourly returns.

This financial-instrument/Twitter-Filter combination is the only one seen in the study for which hourly changes in Twitter sentiments carried no ability to lead an asset's hourly returns, whilst hourly changes in the Twitter message volumes were able to lead the asset's hourly returns in a statistically-significant manner.

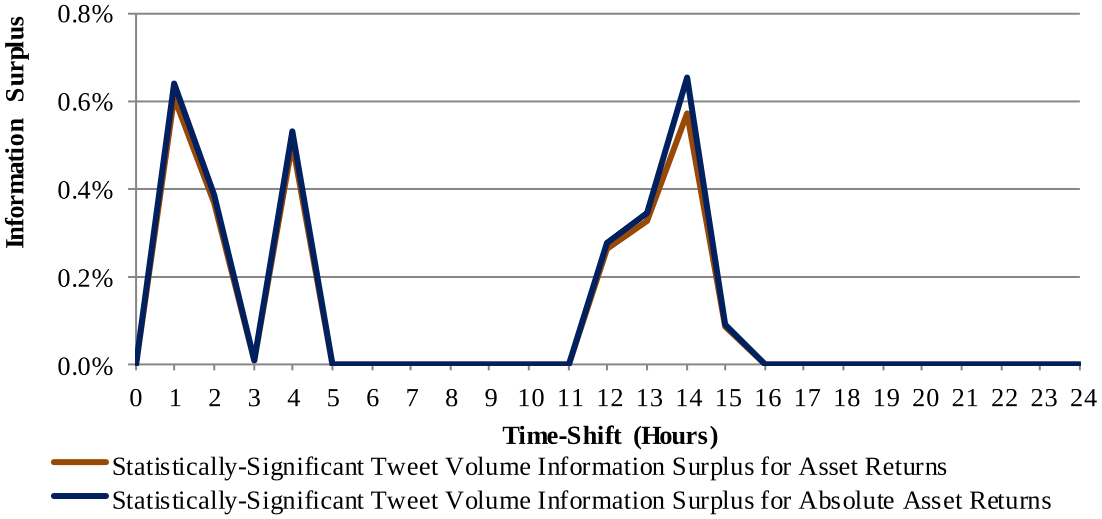


FIGURE 31: TIME-SHIFTS AT WHICH HOURLY CHANGES IN TWEET VOLUMES FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF BANK OF AMERICA, CORP. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

11.1.5 Cisco Systems, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Cisco Systems, Inc. is a networking and communication device business headquartered in California, USA and is listed in the NASDAQ 100 stock index, the AMEX internet index and the Dow Jones index with a market capitalisation of $117bn as at December 2014[a], and is the world's 58th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the design and manufacture of digital networking and communication equipment for both consumer and business customers.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$CSCO" AND/OR "Cisco" (one of Cisco Systems' trading names), to capture Tweets mentioning Cisco Systems, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 537 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Cisco Systems, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Twitter sentiment led asset's hourly returns in a statistically-significant manner. As seen in Figure 32 below, hourly changes in Twitter's negative sentiments on the company were able to lead the hourly returns of Cisco Systems, Inc. CFDs for eight time-shifts, with a peak information surplus of 1.90% occurring at a leading time-shift of 11-hours. Hourly changes in Twitter's positive sentiments on the company were not able to lead the hourly returns of its CFDs on any occasion. Finally, hourly changes in Twitter's net sentiments on the company were able to lead the hourly returns of its CFDs for seven time-shifts, with a peak information surplus of 2.77% at a leading time-shift of 13-hours, indicating that hourly changes in Twitter's net sentiment on Cisco Systems are most indicative of the hourly returns of the asset's CFDs ahead of time.

[a] http://uk.finance.yahoo.com/q/pr?s=CSCO

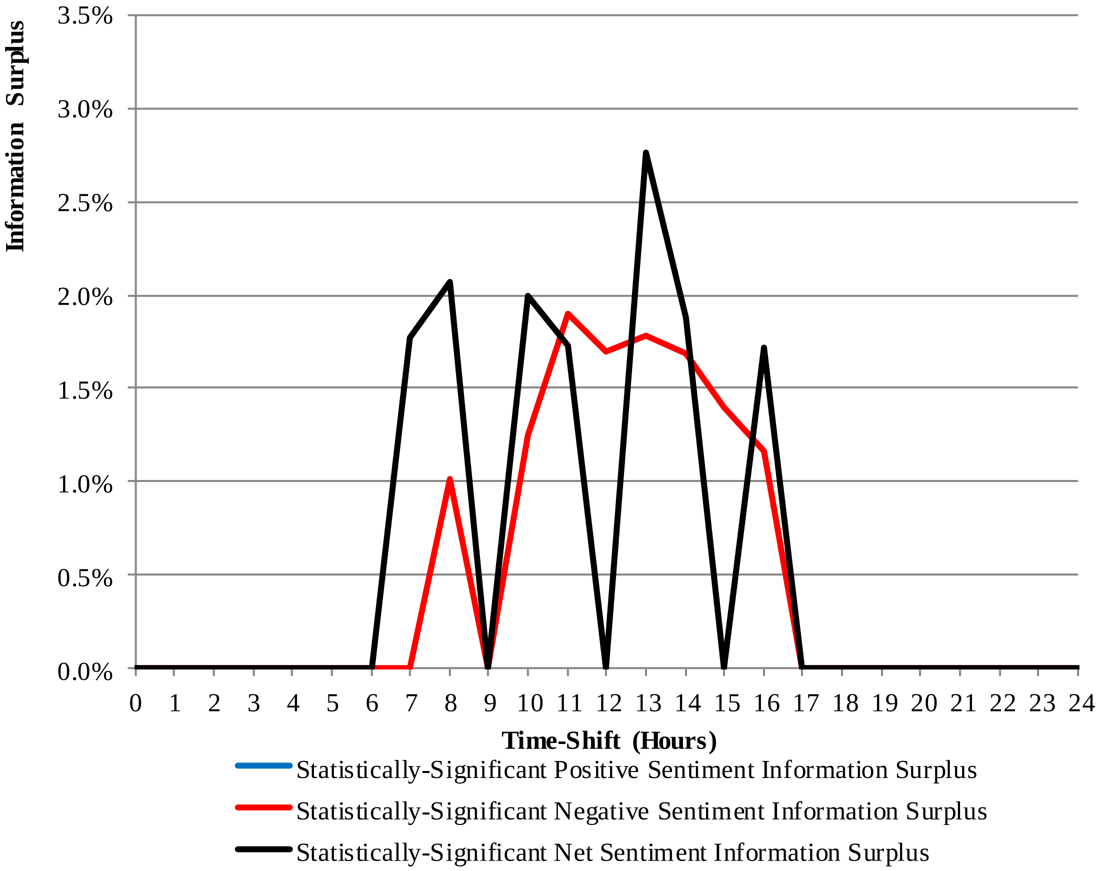


FIGURE 32: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF CISCO SYSTEMS, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 32 shows the range of statistically-significant information surplus values for the three sentiment types.

Hourly changes in Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of Cisco Systems, Inc. CFDs in a statistically-significant manner.

11.1.6 Google, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Google, Inc. is an internet information provider headquartered in California, USA and is listed in the NASDAQ 100 stock index and the AMEX internet index with a market capitalisation of $342bn as at December 2014[a], and is the world's 2nd highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the development of online-based services for organising and searching through information, with a particular emphasis on internet search.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$GOOG" AND/OR "Google", to capture Tweets mentioning Google, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 24.5 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Google, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 33 below, hourly changes in Twitter's negative sentiments on the company were able to lead the hourly returns of Google, Inc. CFDs for three time-shifts, with a peak information surplus of 0.96% occurring at a leading time-shift of 2-hours. Hourly changes in Twitter's positive sentiments on the company were not able to lead the hourly returns of its CFDs on any occasion. Finally, hourly changes in Twitter's net sentiments on the company were able to lead the hourly returns of its CFDs for eleven time-shifts, with a peak information surplus of 2.63% at a leading time-shift of 14-hours, indicating that hourly changes in Twitter's net sentiment on Google is most indicative of the hourly returns of the asset's CFDs ahead of time.

[a] http://uk.finance.yahoo.com/q/pr?s=GOOG

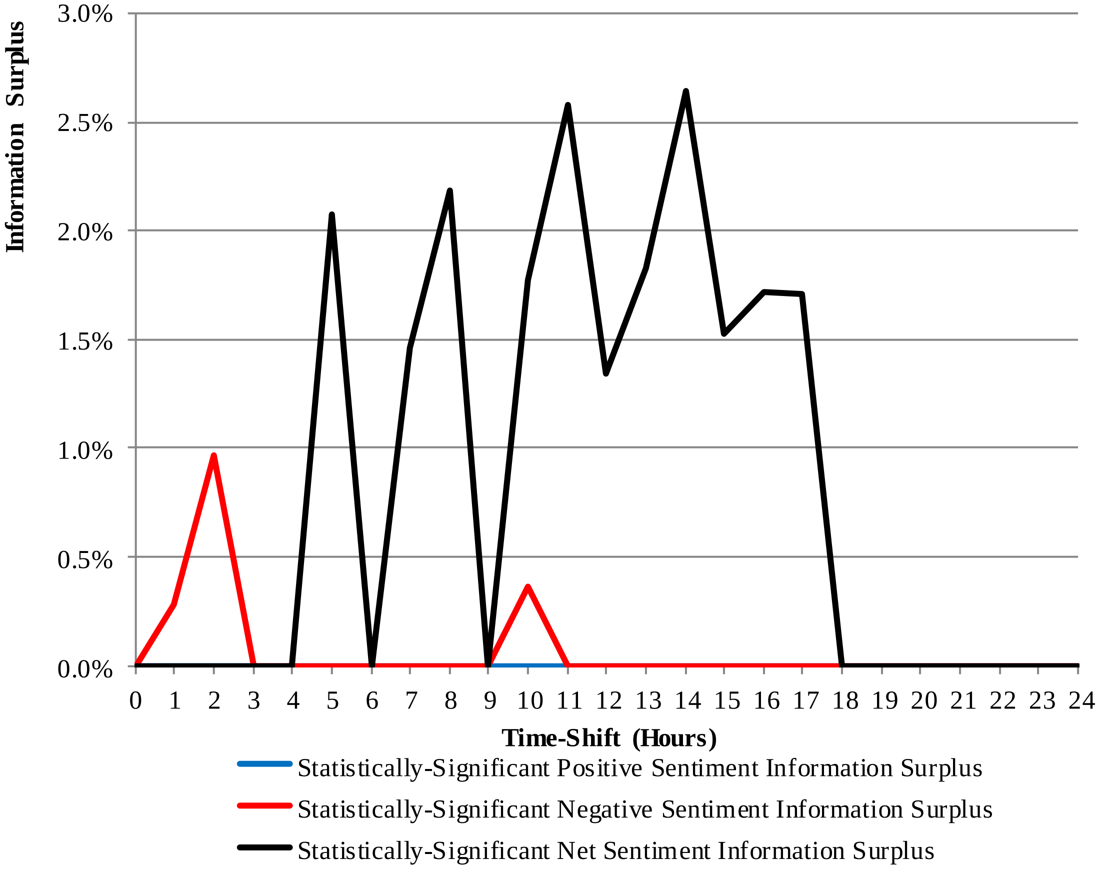


FIGURE 33: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF GOOGLE, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 33 shows the range of statistically-significant information surplus values for the three sentiment types.

Hourly changes in Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of Google, Inc. CFDs in a statistically-significant manner.

11.1.7 The Home Depot, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

The Home Depot, Inc. is a leading operator of home-improvement retail stores headquartered in Georgia, USA and is listed in the Dow Jones Composite stock index with a market capitalisation of $130bn as at December 2014[a], and is the world's 18th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of building materials, equipment and services for consumer-centric home-improvement and home-maintenance purposes.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$HD" AND/OR "Home Depot" (one of The Home Depot's trading names), to capture Tweets mentioning The Home Depot, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 251 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of The Home Depot, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 34 below, hourly changes in Twitter's positive sentiments on the company were able to lead the hourly returns of The Home Depot, Inc. CFDs for seven time-shifts, with a peak information surplus of 2.81% occurring at a leading time-shift of 11-hours. Hourly changes in Twitter's negative sentiments on the company were not able to lead the hourly returns of its CFDs on any occasion. Finally, hourly changes in Twitter's net sentiments on the company were able to the hourly lead returns of its CFDs for only one time-shift, with a peak information surplus of 2.28% at a leading time-shift of 9-hours.

[a] http://uk.finance.yahoo.com/q/pr?s=HD

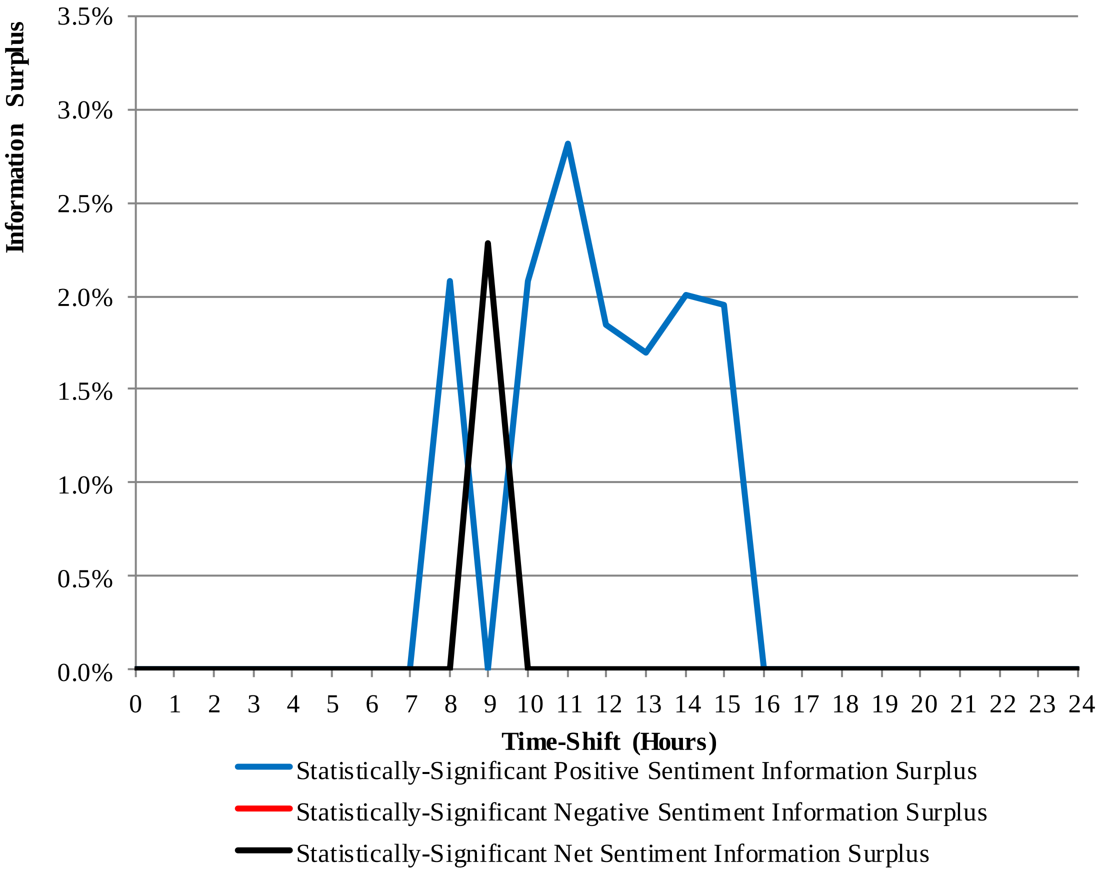


FIGURE 34: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF THE HOME DEPOT, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 34 shows the range of statistically-significant information surplus values for the three sentiment types.

In addition, this financial-instrument/Twitter-Filter combination demonstrated instances of when hourly changes in Tweet message volumes led the asset's hourly returns ahead of time in a statistically-significant manner. As seen in Figure 35 below, hourly changes in Tweet message volumes led the hourly returns of The Home Depot, Inc. CFDs for four time-shifts, with a peak information surplus of 2.02% occurring at a time-shift of 15-hours. Hourly changes in Tweet message volumes led the absolute hourly returns of The Home Depot, Inc. CFDs also for four time-shifts, with a peak information surplus of 2.23% occurring at a leading time-shift of 15-hours. This indicates that hourly changes in Tweet message volumes show a greater capacity to lead the asset's absolute hourly returns than the asset's actual hourly returns.

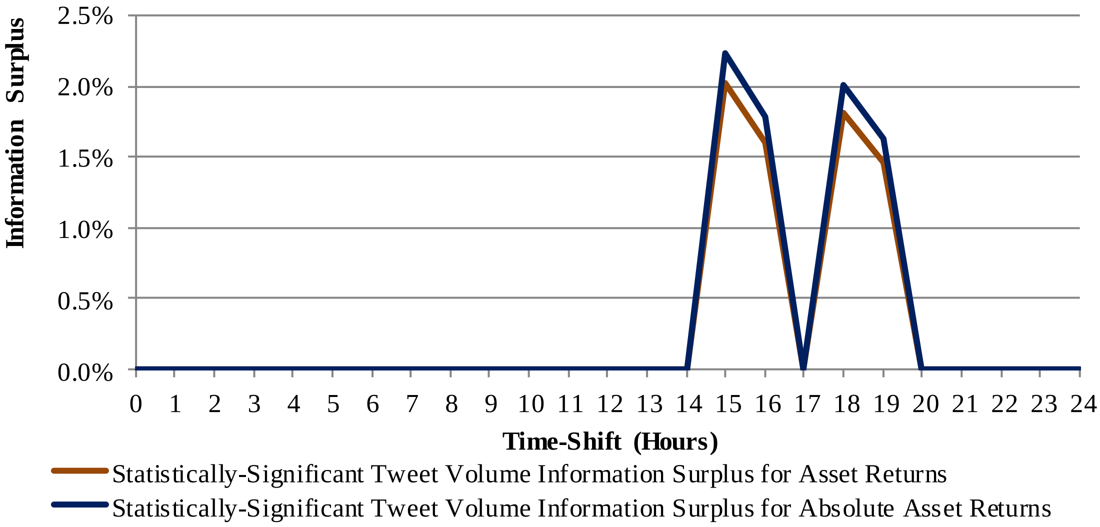


FIGURE 35: TIME-SHIFTS AT WHICH HOURLY CHANGES IN TWEET VOLUMES FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF THE HOME DEPOT, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

11.1.8 Intel, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Intel, Corp. is a designer and manufacturer of integrated digital technology headquartered in California, USA and is listed in the NASDAQ 100 index and the Dow Jones Composite stock index with a market capitalisation of $175bn as at December 2014[a], and is the world's 26th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the business-to-business provision of microprocessor and chipset hardware for ultimate use in mobile, professional and consumer computing applications.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$INTC" AND/OR "Intel", to capture Tweets mentioning Intel, Corp.'s Ticker-ID AND/OR the name of the company. In this manner, 1.7 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the returns of Intel Corp. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated only two time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 36 below, hourly changes in Twitter's negative sentiments on the company were able to lead the hourly returns of Intel, Corp. CFDs for these two time-shifts, with a peak information surplus of 1.41% occurring at a leading time-shift of 1-hour. Hourly changes in neither Twitter's positive sentiments on the company, nor the net sentiments, were able to lead the hourly returns of its CFDs on any occasion.

[a] http://uk.finance.yahoo.com/q/pr?s=INTC

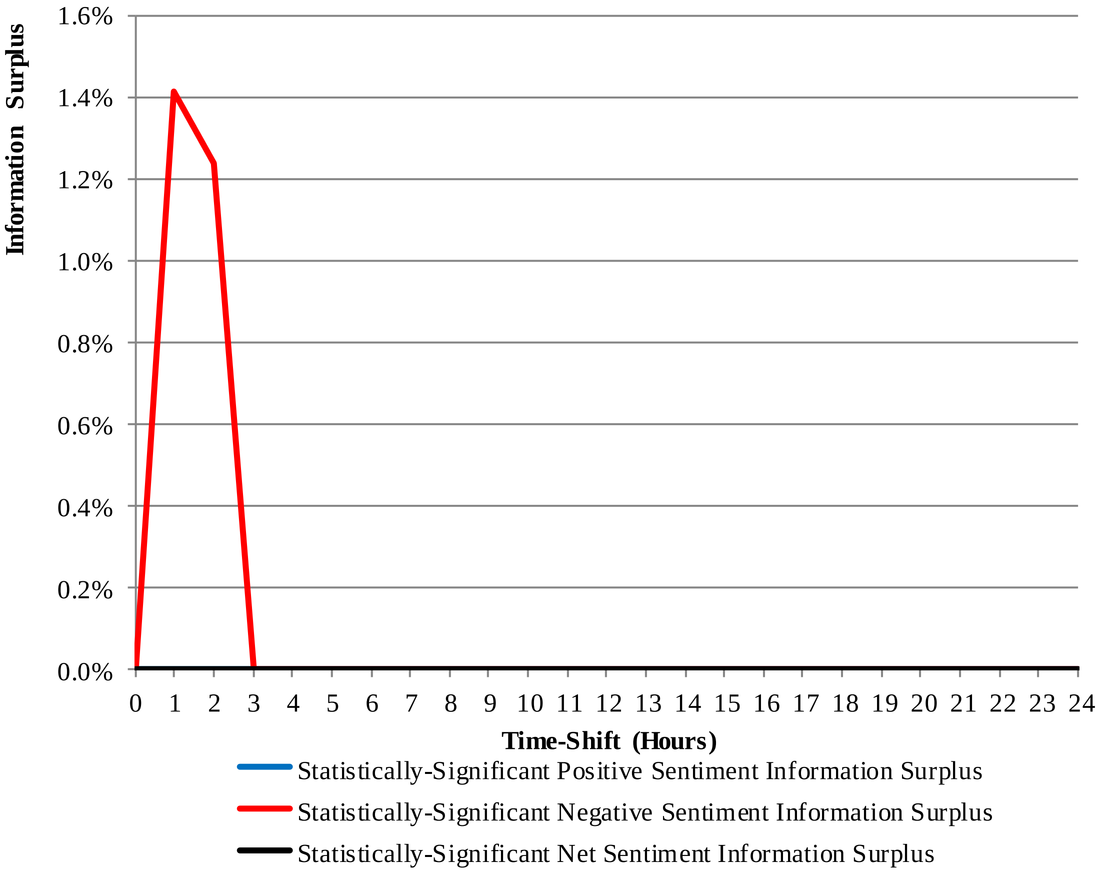


Figure 36: Time-shifts at which hourly changes in sentiments of Tweets filtered by Ticker-ID AND/OR Company Name led the hourly returns of Intel, Corp. CFDs in a statistically-significant manner

Figure 36 shows the small range of statistically-significant information surplus values for the three sentiment types, with only hourly changes in the negative sentiment type being able to lead the asset’s hourly returns, with a peak information surplus of 1.41% at a leading time-shift of 1-hour.

In addition, this financial-instrument/Twitter-Filter combination demonstrated instances of when hourly changes in Tweet message volumes was able to lead the asset’s hourly returns ahead of time in a statistically-significant manner. As seen in Figure 37 below, hourly changes in Tweet message volumes led the absolute hourly returns of Intel, Corp. CFDs on two occasions, with a peak information surplus of 0.52% occurring at a time-shift of 2-hours.

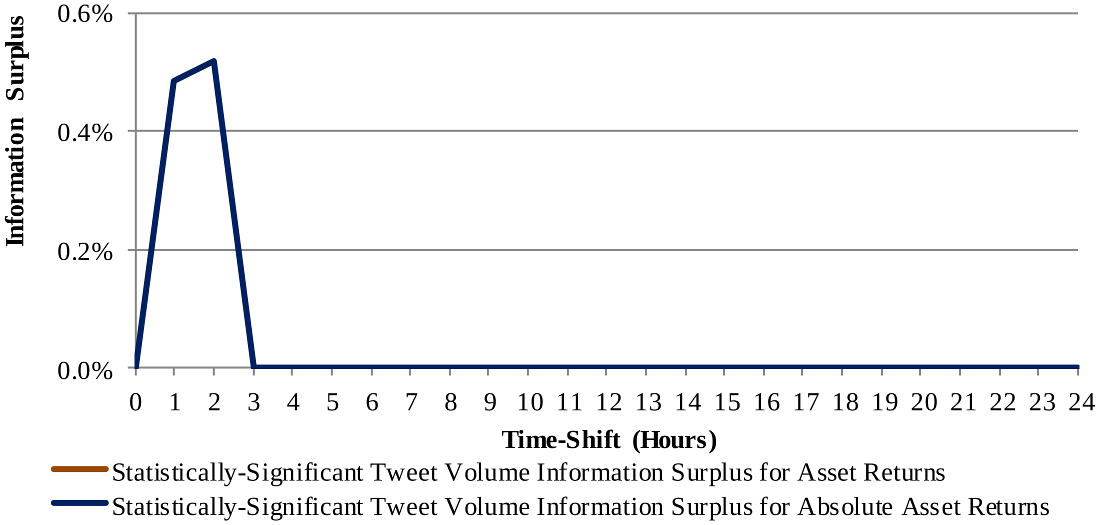


Figure 37: Time-shifts at which hourly changes in Tweet volumes filtered by Ticker-ID AND/OR Company Name led the hourly returns of Intel, Corp. CFDs in a statistically-significant manner

11.1.9 J.P. Morgan, Inc. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

J.P. Morgan, Inc. is a financial services provider and international bank headquartered in New York, USA and is listed in the Dow Jones Composite stock index with a market capitalisation of $223bn as at December 2014[a], and is the world's 65th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of commercial banking services such as Mergers & Acquisitions, Initial Public Offerings, market-making, asset brokerage and commercial debt finance.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$JPM" AND/OR "JPMorgan" AND/OR "JP Morgan" (J.P. Morgan's trading names), to capture Tweets mentioning J.P. Morgan, Inc.'s Ticker-ID AND/OR the name of the company. In this manner, 133 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of J.P. Morgan, Inc. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated only two time-shifts for which hourly changes in Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 38 below, hourly changes in Twitter's positive sentiments on the company were able to lead the returns of J.P. Morgan, Inc. CFDs for these two time-shifts, with a peak information surplus of 3.93% occurring at a leading time-shift of 12-hours. Hourly changes in neither Twitter's negative sentiments on the company, nor the net sentiments, were able to lead the hourly returns of its CFDs on any occasion.

[a] http://uk.finance.yahoo.com/q/pr?s=JPM

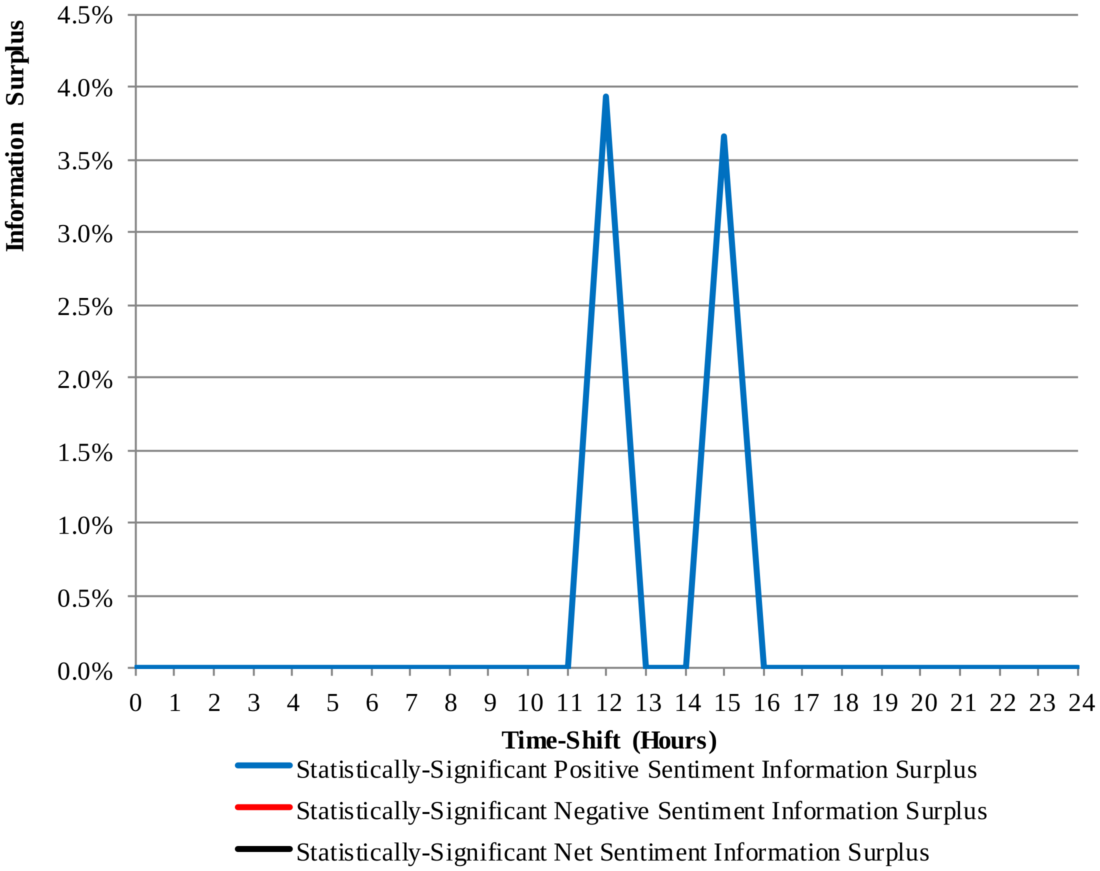


FIGURE 38: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF J.P. MORGAN, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 38 shows the small range of statistically-significant information surplus values for the three sentiment types, with only hourly changes in the positive sentiment type being able to lead the asset's hourly returns. Here, a peak information surplus of 3.93% was seen at a leading time-shift of 12-hours.

In addition, this financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in Tweet message volumes led the asset's hourly returns ahead of time in a statistically-significant manner. As seen in Figure 39 below, hourly changes in Tweet message volumes led the hourly returns of J.P. Morgan, Inc. CFDs for fifteen time-shifts, with a peak information surplus of 1.21% occurring at a time-shift of 13-hours. Hourly changes in Tweet message volumes led the absolute hourly returns of J.P. Morgan, Inc. CFDs for fourteen time-shifts, with a peak information surplus of 1.37% occurring at a leading time-shift of 16-hours. This indicates that hourly changes in Tweet message volumes show a greater capacity to lead the asset's absolute hourly returns than the asset's actual hourly returns.

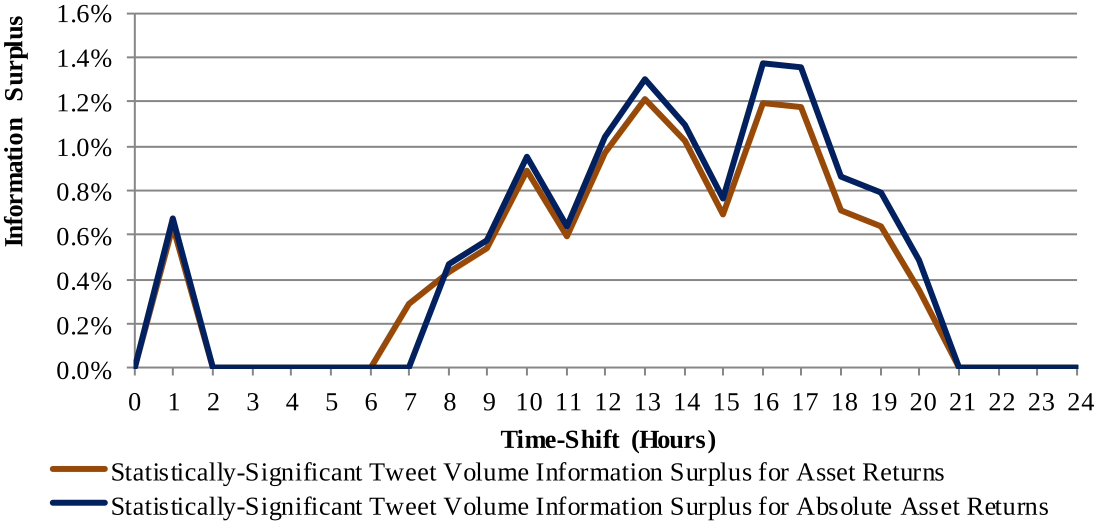


FIGURE 39: TIME-SHIFTS AT WHICH HOURLY CHANGES IN TWEET VOLUMES FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF J.P. MORGAN, INC. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

11.1.10 Coca-Cola, Co. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Coca-Cola, Co. is a beverage manufacturer and distributor headquartered in Georgia, USA and is listed in the Dow Jones Composite stock index with a market capitalisation of $182bn as at December 2014[a], and is the world's 9th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the worldwide manufacture, marketing and sale of non-alcoholic beverages.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$KO" AND/OR "Coca Cola" AND/OR "Coca-Cola" (Coca-Cola's trading names), to capture Tweets mentioning Coca-Cola, Co.'s Ticker-ID AND/OR the name of the company. In this manner, 3.3 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Coca-Cola, Co. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in the Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 40 below, hourly changes in the Twitter's positive sentiments on the company were able to lead the hourly returns of Coca-Cola, Co. CFDs for thirteen time-shifts, with a peak information surplus of 0.72% occurring at a leading time-shift of 8-hours. Hourly changes in the Twitter's net sentiments on the company were not able to lead the hourly returns of its CFDs on any occasion. Finally, hourly changes in the Twitter's negative sentiments on the company were able the hourly returns of its CFDs for just one time-shift, with a peak information surplus of 0.06% at a leading time-shift of 1-hours, indicating that hourly changes in the Twitter's positive sentiment on Coca-Cola is most indicative of the returns of the asset's CFDs ahead of time.

[a] http://uk.finance.yahoo.com/q/pr?s=KO

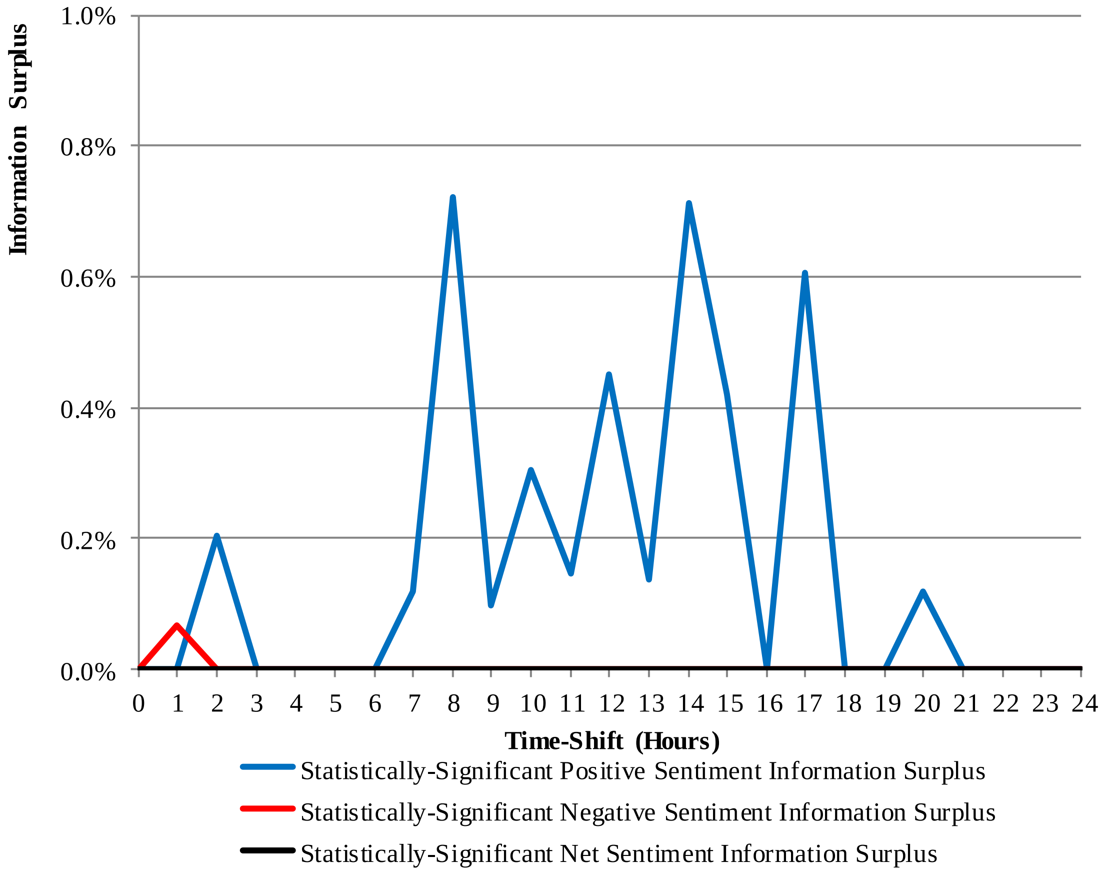


Figure 40: Time-shifts at which hourly changes in sentiments of Tweets filtered by Ticker-ID AND/OR Company Name led the hourly returns of Coca-Cola, Co. CFDs in a statistically-significant manner

Figure 40 shows the range of statistically-significant information surplus values for the three sentiment types.

Hourly changes in the Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of Coca-Cola, Co. CFDs in a statistically-significant manner.

11.1.11 McDonald's, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

McDonald's, Corp. is an operator and franchiser of fast-food restaurants headquartered in Illinois, USA and is listed in the Dow Jones Composite stock index with a market capitalisation of $90bn as at December 2014[a], and is the world's 25th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of fast-food outlets under the McDonald's name.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$MCD" AND/OR "McDonald's", to capture Tweets mentioning McDonald's, Corp.'s Ticker-ID AND/OR the name of the company. In this manner, 6.1 million Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of McDonald's, Corp. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated multiple time-shifts for which hourly changes in the Twitter sentiment led the asset's hourly in a statistically-significant manner. As seen in Figure 41 below, hourly changes in the Twitter's net sentiments on the company were able to lead the hourly returns of McDonald's, Corp. CFDs for six time-shifts, with a peak information surplus of 1.90% occurring at a leading time-shift of 13-hours. Hourly changes in the Twitter's positive sentiments on the company were not able to lead the hourly returns of its CFDs on any occasion. Finally, hourly changes in the Twitter's negative sentiments on the company were able to lead the hourly returns of its CFDs for just one time-shift, with a peak information surplus of 1.28% at a leading time-shift of 7-hours, indicating that Twitter's net sentiment on McDonald's is most indicative of the returns of the asset's CFDs ahead of time.

[a] http://uk.finance.yahoo.com/q/pr?s=MCD

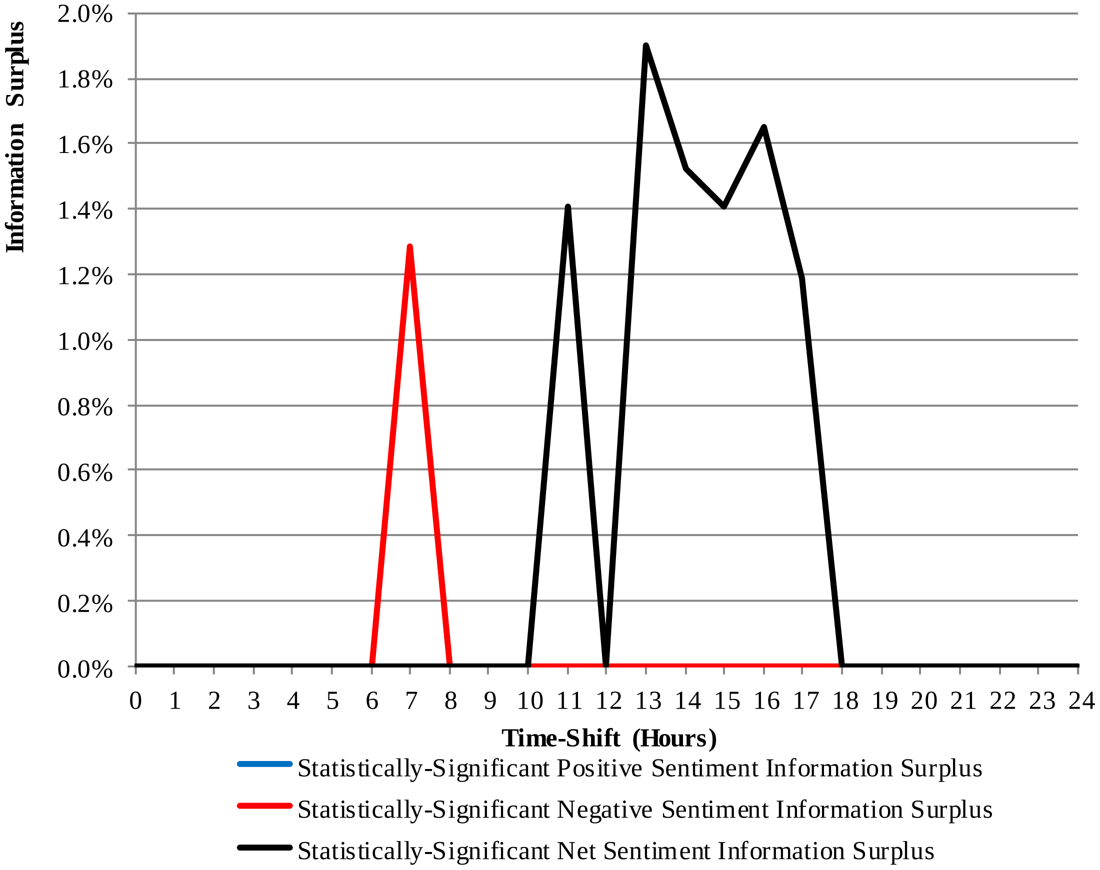


FIGURE 41: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF MCDONALD'S, CORP. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 41 shows the range of statistically-significant information surplus values for the three sentiment types, with only hourly changes in the negative and net sentiment types being able to lead the asset’s hourly returns.

Hourly changes in the Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of McDonald’s, Corp. CFDs in a statistically-significant manner.

11.1.12 Oracle, Corp. CFDs, with social media data filtered by Ticker-ID AND/OR Company Name

Oracle, Corp. is a developer and manufacturer of database and middleware software headquartered in California, USA and is listed on the New York Stock Exchange with a market capitalisation of $182bn as at December 2014[a], and is the world's 55th highest-ranking company by brand popularity[111] at the time of writing. It is primarily involved in the provision of electronic database management services to corporate, rather than retail customers.

For this financial-instrument/Twitter-Filter combination, Tweets were filtered from Twitter's 10% Gardenhose Feed without any geographical filtering using the string-filter: "$ORCL" AND/OR "Oracle", to capture Tweets mentioning Oracle's Ticker-ID AND/OR the name of the company. In this manner, 654 thousand Tweets were filtered in during this study's 3-month data-collection period, and subsequently analysed to ascertain the extent to which they can lead the hourly returns of Oracle, Corp. CFDs.

This financial-instrument/Twitter-Filter combination demonstrated just one time-shift for which hourly changes in the Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 42 below, hourly changes in the Twitter's net sentiments on the company were able to lead the hourly returns of Oracle, Corp. CFDs for this one time-shift, with a peak information surplus of 0.36% occurring at a leading time-shift of 1-hours. Hourly changes in the neither Twitter's positive sentiments on the company, nor the negative sentiments, were able to lead the hourly returns of its CFDs on any occasion.

[a] http://uk.finance.yahoo.com/q/pr?s=ORCL

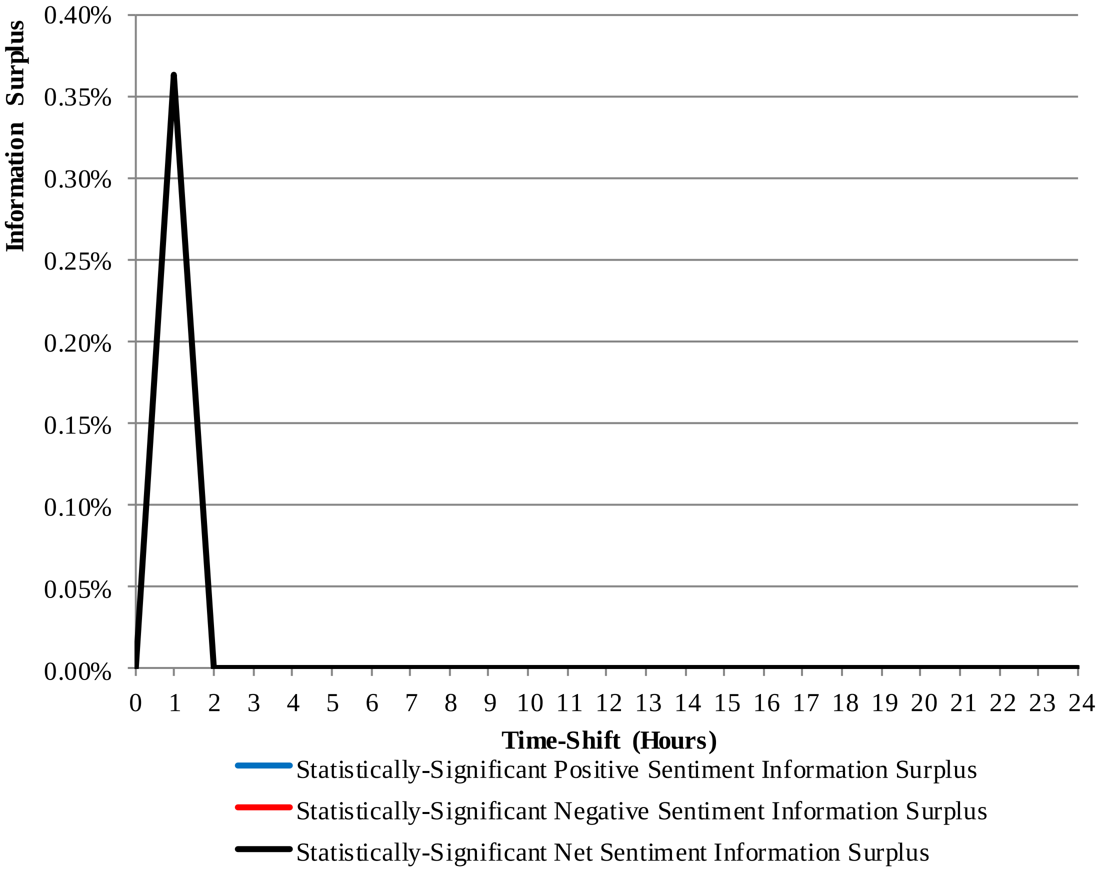


FIGURE 42: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF TWEETS FILTERED BY TICKER-ID AND/OR COMPANY NAME LED THE HOURLY RETURNS OF ORACLE, CORP. CFDS IN A STATISTICALLY-SIGNIFICANT MANNER

Figure 42 shows the limited range of statistically-significant information surplus values for the three sentiment types, with only hourly changes in the net sentiment type being able to lead the asset's hourly returns.

Hourly changes in the Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of Oracle, Corp. CFDs in a statistically-significant manner.

11.1.13 S&P500 Index Futures, with social media data sourced from string-unfiltered Tweets of US-origin

The same collection process was used for this financial-instrument/Twitter-Filter combination as the S&P500 Index CFDs (Chapter 6.3.1.6). Thus, 18.7 million Tweets analysed to ascertain the extent to which they can lead the hourly returns of S&P500 Index Futures.

This financial-instrument/Twitter-Filter combination demonstrated just one time-shift for which hourly changes in the Twitter sentiment led the asset's hourly returns in a statistically-significant manner. As seen in Figure 43 below, hourly changes in the Twitter's net sentiments from the US were able to lead the hourly returns of the S&P500 Index Futures for this one time-shift, with a peak information surplus of 2.46% occurring at a leading time-shift of 22-hours. Hourly changes in the neither Twitter's positive sentiments from the US, nor the negative sentiments, were able to lead the hourly returns of S&P500 Index Futures on any occasion.

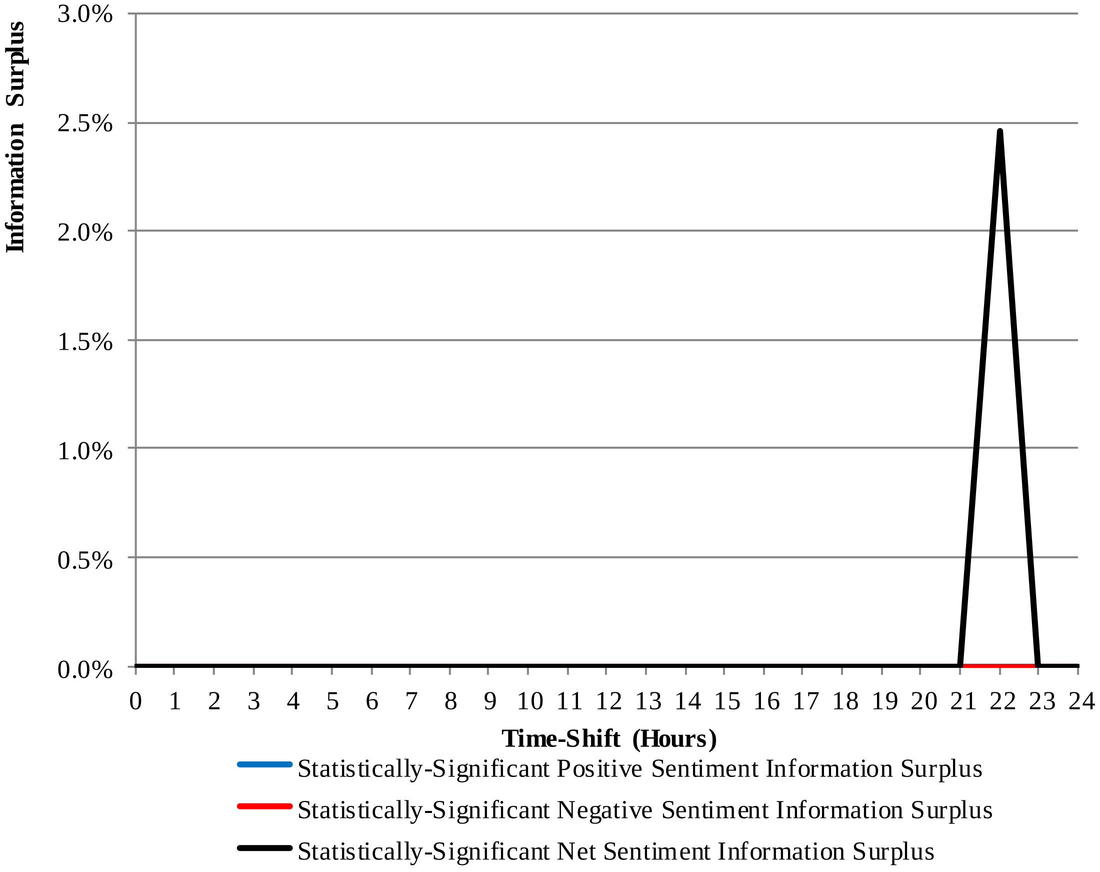


FIGURE 43: TIME-SHIFTS AT WHICH HOURLY CHANGES IN SENTIMENTS OF STRING-UNFILTERED TWEETS FROM THE US LED THE HOURLY RETURNS OF S&P500 INDEX FUTURES IN A STATISTICALLY-SIGNIFICANT MANNER

Hourly changes in the Tweet message volumes produced using this Twitter filter showed no ability to lead the hourly returns or the absolute hourly returns of S&P500 Index Futures in a statistically-significant manner.

11.2 Code and raw data Appendix

The HTML link below contains access to:

- Readable and runnable code for the Java-based Twitter Collection Framework (TCF);
- Readable and runnable code for the MATLAB-based Statistical analysis Framework (SAF);
- Readable and runnable code for the MATLAB-based Time Series Processing Framework (TSPF);
- The manual for SocialSTORM;
- The Licence Agreement with Twitter providing access to the network's 10% Gardenhose Feed;
- The raw price data used in this study;
- The raw Twitter data used in this study.

LINK:
http://goo.gl/1Po2h8