ANALYSIS OF NEWS SENTIMENT AND ITS APPLICATION TO FINANCE

By

Xiang Yu

A thesis submitted for the degree of Doctor of Philosophy

School of Information Systems, Computing and
Mathematics,
Brunel University

6 May 2014

Dedication: To the loving memory of my mother.

Abstract

We report our investigation of how news stories influence the behaviour of tradable financial assets, in particular, equities. We consider the established methods of turning news events into a quantifiable measure and explore the models which connect these measures to financial decision making and risk control.

The study of our thesis is built around two practical, as well as, research problems which are determining trading strategies and quantifying trading risk. We have constructed a new measure which takes into consideration (i) the volume of news and (ii) the decaying effect of news sentiment. In this way we derive the impact of aggregated news events for a given asset; we have defined this as the impact score. We also characterise the behaviour of assets using three parameters, which are return, volatility and liquidity, and construct predictive models which incorporate impact scores.

The derivation of the impact measure and the characterisation of asset behaviour by introducing liquidity are two innovations reported in this thesis and are claimed to be contributions to knowledge.

The impact of news on asset behaviour is explored using two sets of predictive models: the univariate models and the multivariate models. In our univariate predictive models, a universe of 53 assets were considered in order to justify the relationship of news and assets across 9 different sectors. For the multivariate case, we have selected 5 stocks from the financial sector only as this is relevant for the purpose of constructing trading strategies. We have analysed the celebrated Black-Litterman model (1991) and constructed our Bayesian multivariate predictive models such that we can incorporate domain expertise to improve the predictions. Not only does this suggest one of the best ways to choose priors in Bayesian inference for financial models using news sentiment, but it also allows the use of current and synchronised data with market information. This is also a novel aspect of our work and a further contribution to knowledge.

Acknowledgements

First and foremost, I thank my supervisors Prof. Gautam Mitra and Prof. XiaoHui Liu for their continuous support and guidance during these three and a half years. It has been the most comforting thought to know that they are here to help me.

I am grateful for the financial support provided to me by the Engineering and Physical Sciences Research Council (EPSRC) and OptiRisk Systems.

The Department of Mathematical Sciences at Brunel University has offered me an excellent environment in which to carry out my research and more importantly given me the opportunity to meet colleagues who have become my close friends.

Most importantly, I thank my family for their unconditional love and support, especially my father for directing me along this educational path.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

Although most of the abbreviations are explained when they are used for the first time in the text, we also list them here.

ADF	Augmented Dickey-Fuller test
AR	Autoregressive
BVAR	Bayesian vector autoregressive model
CAPM	Capital asset pricing model
DJIA	Dow Jones Industrial average
EMH	Efficient market hypothesis
ETF	Exchange-traded fund
GARCH	Generalised autoregressive conditional heteroskedasticity
GI	General Inquirer of the Harvard Dictionary
HFT	High Frequency Trading
MAE	Mean absolute error
MCMC	Markov Chain Monte Carlo
MEC	Market-efficient coefficient
MGARCH	Multivariate GARCH
PIN	Probability of informed trading
RMSE	Root mean squared error
TED Spread	TED spread is the difference in interest rates on interbank loans and short-term U.S. government debt subsequently forming the acronym for T-Bill and ED (the ticker symbol for Eurodollar futures contract).
VAR	Vector autoregressive
VIX	Volatility implied index

Chapter 1

Introduction

1.1 Focus of the Thesis

The research represented in this thesis focuses on exploring news as an event and the incorporation of news in prediction models to enhance the power of predictability. Specifically the sentiment of news is taken as a quantitative reflection of information. Methods of classifying news sentiment and the conversion process of machine-readable text to a quantitative measure are considered; the main focus of our work is to study the applications of news sentiment in a financial context. Particular attention is directed towards higher frequencies of trading, i.e. minute bar, due to the greater profitability prospects and higher number of trading signals related to news in these time frequencies. Since the most popular trading instrument in high frequencies is also the most reported type of securities in news, we have selected this as the asset class of study, namely equities.

1.2 News as an Event

"**What you see is news, what you know is background, what you feel is opinion**."
Lester Markel, American journalist, 1894-1977

News is publicly shared information that is accessible to the mass audience. In the domain of Finance, news is reported on companies, stocks, regulations and more. News usually covers an event and is dispersed as noteworthy facts to share information and gain knowledge about the markets. A consequence of news distribution is that recipients of the news form opinions and enable investors to make judgements, which could lead to actions being taken in their portfolio of assets. The subsequent reaction of the market participants due to sentiment formed from a piece of news is what we are interested in studying in this thesis.

The steps preceding the public dissemination of news are crucial in sustaining the validity and credibility of news stories. The process involves information gathering through credible sources so that journalists are able to produce a written article which then goes through editorial control. The scrutiny and oversight of this editorial process ensures that only the validated items of information are published to maintain readers' confidence in the trustworthiness of news. Furthermore, this also implies that a process of filtering occurs to determine whether a piece of news is adequate for broadcasting. The reasoning can be explained as follows. The information is exclusively known to a small group of individuals to begin with, which is subsequently passed on to journalists who decide whether it is necessary to be publicised. Thus a selection criterion is imposed to differentiate between news that is suitable for distribution and those that should remain private. In other words, news is a class of information that is deemed noteworthy for market participants (investors, traders and more) to see. Naturally this raises the question of who controls such a selection. The intuitive answer would be journalists; upon further contemplation it becomes apparent that it is the source who provides the journalist with such information. Hence, there must be some form of motivation driving the source to release it and require it to be shared publicly. Manipulation of such a process occurs when false or misleading information is distributed with the aim of moving markets to produce financial gain for the person. This is known as market manipulation and is illegal. For journalists, their objective is to attract readers' attention and create interest, which may lead to the formation of a common opinion amongst market participants. This is loosely formed as sentiment.

Given that news stories are often repeated it follows that not all news stories may be novel and hence informative. If news is rapidly digested by markets then the efficient market hypothesis (EMH) holds meaning it is not possible to achieve returns larger than average market returns. Indeed this has been the foundation of finance theory for decades, ever since its proponent and now a Nobel Laureate (2013) Professor Eugene Fama developed the hypothesis in 1965 as part of his PhD thesis. The EMH claims that asset prices already reflect all past publicly available information, and change instantly to reflect new public information, even insiders' information. There are three levels of EMH: weak, semi-strong and strong. The neoclassical finance axiom of all traders being rational is guided by utility theory. However, as with all landmark theories, there are many studies contesting this frame of thought such as Rosenberg, Reid, and Lanstein (1985) and Shiller et al. (1984). Additionally, Nicholson (1968) and Basu (1977) find a group of stocks that would outperform others in the form of low price to earnings ratio stocks, violating EMH. By the 1990s, behavioural finance began to be widely accepted, bringing in the opposition – irrational markets (Kahneman, 2002, Shiller, 2000 and Shefrin, 2008). Postulated by prospect theory (see Kahneman and Tversky, 1979) the argument for irrational traders emerged. Evidence was compiled for over reactions and under reactions to new information, causing excess volatility to appear in the markets. At that time, an explanation for these findings was that they were simply reflecting a complex pattern of compensation for systematic risk, which in classical finance theory is the only determining factor of expected returns. The criticism of EMH is now well accepted and is termed “market anomalies”. Some critics blame the blind faith and pursuit of EMH as the underlying reason for the recent global financial crisis (Nocera 5 June 2009).

A finer investigation into the markets exposes cases where inefficient behaviour is prevalent. Intuitively, small stocks react to news significantly slower than large stocks, with respect to their market capitalisation. This can be attributed to many factors such as illiquidity of their stocks, due to the lack of awareness about them, in addition to a delay in reaction of market prices also caused by the same reason plus a lack of news events. Baker and Wurgler (2006) report similar findings in their work and state that small stocks earn particularly high returns in low sentiment periods, but when sentiment is high there is no size effect. On the other hand, a single news item has a significantly larger effect on small stocks than big companies as news is rarely reported therefore, any news provides more insight into such a company. Barber and Odean (2008) declare any unusual trading volume in large capitalisation stocks is unlikely to be driven by only a few investors, as is the case for smaller and more illiquid stocks. Baker and Wurgler (2006) argue the case for irrational markets by presenting evidence of investor sentiment affecting stock prices, rejecting the neoclassical finance theory that all markets are rational markets. In their paper, sentiment is measured by a proxy - closed-end fund discount (CEFD), which is calculated as the average difference between the net asset values of closed-end stock fund shares and their market prices. They suggest that this measure is inversely related to sentiment. Empirical results find that cross-sectional predictability patterns in stock returns are conditional upon proxies for beginning-of-period sentiment.

Until recently, the concept of sentiment, especially news sentiment, had still been a very elusive concept. The common school of thought was that sentiment offsets itself between the opposing opinions. However, a collection of research can be compiled to argue that text contained in news stories is important; they influence and move the market. Fisher and Statman (2000) recognise sentiment as an important component of the market pricing process by surveying a panel of individual investors, newsletter writers and Wall Street sell-side analysts. Chan (2003) investigates the difference in behaviour between assets that are reported in the news and assets which are not covered. He finds a strong drift in returns after bad news is released whereas price movements of assets unaccompanied by news experience annual return reversals. Moreover, Barber and Odean (2008) and Sinha's (2010) perspective is that news releases and market price updates occur at different paces hence a causal relationship exists. Alternatively, investors' sentiment can also be considered as a reflection of their perception on risk. For example, in periods where markets are bearish and confidence is low, only news with high positive sentiment would be sufficient to persuade traders to take drastic actions, indicating their risk adverse behaviour. Theory suggests that firms with certain characteristics are most volatile to shifts in sentiment, namely, newer, smaller, more volatile, unprofitable, non-dividend paying, distressed or extreme growth potential firms (Baker and Wurgler, 2006).

Subsequently, research literature in this field of news sentiment analysis has grown dramatically over the past few decades. The main focus of work has been on investigating the role of sentiment in explaining the time series of returns (Kothari and Shanken, 1997, Neal and Wheatley, 1998, Shiller, 2000, Baker and Wurgler, 2000). In particular, the work of DeLong, Shleifer, Summers and Waldmann (1990) needs to be especially highlighted as they were one of the first to claim that securities with more exposure to sentiment have higher unconditional expected returns. Only in recent years has a body of literature emerged studying the relationship and effects of sentiment on stock market variables other than returns, for example, volatility and liquidity (Gross-Klaussman and Hautsch, 2011, Smales, 2013, Riordan et al., 2013).

Taken from market microstructure theory, traders can also be categorised according to the information they know. There exist two types of traders, where one party (informed traders) possesses more information about market movements than its counterpart (uninformed traders). Institutional traders constitute the informed party whereas individual traders compose the uninformed party, also known as noise traders.

The behaviour of informed and uninformed traders differs substantially (see Barber and Odean, 2008 and Tetlock, 2011). Uninformed traders, such as retail traders, have been proven to react to news events immediately after announcement whereas institutional traders do not.

1.3 Sentiment and Its Evolution

The definition of "sentiment" has evolved as research progressed. The initial sentiment that was of interest to researchers was the views and opinions of investors specifically i.e. **investor sentiment** (DeLong et al., 1990). Whether it is their direct response to an announcement or after reading a news article, there was motivation to suggest that stock price movements would be related to such sentiment. Often these opinions do not translate into instant reactions but rather evaluated over time so that decisions could be made for the long term. Naturally a reflection of the mass investors' views is the media, which gives a summary of the general mood and consensus of the stock markets. Henceforth studies tried to capture the views known as **media sentiment** (Tetlock, 2007). This is derived directly from the text of news stories and simply analyses the number of positive and negative words to deduce the overall sentiment. Eventually this was refined to news sentiment (Barber and Odean, 2008, Sinha, 2011, Smales, 2012), which is what we study in this thesis. The classification methods of such sentiment are now more sophisticated and discussed in further detail in section 1.4. It is worth mentioning the final form of sentiment, although it is not reported in our review, that is, **market sentiment**. While lacking in a precise definition, this type of sentiment includes more than just news information; market conditions and VIX are amongst the other factors considered. It is dominated by the consensus of informed traders, who receive and digest extra information regarding future prices, and so contrasts with the underlying assumption of EMH that all markets efficiently reflect all information. This is particularly the case for the strong form of the hypothesis.

"Never awake me when you have good news to announce, because with good news nothing presses; but when you have bad news, arouse me immediately, for then there is not an instant to be lost."

Napoleon Bonaparte, French Emperor, 1769-1821

Studies have found evidence of a similar concept to this quote in the stock markets, that is, stock prices react more drastically to news that contain negative sentiment as opposed to positive sentiment (Chen et al, 2003, Tetlock, 2007, Barber and Odean, 2008). This effect is exhibited by increased levels of volatility, traded volume and absolute order imbalance in the presence of negative news. Engle and Ng (1993) find an impact of news shocks on the Japanese stock market and reported an asymmetric effect such that negative news has a substantial influence on volatility than positive news. They argue that these results occurr due to the leverage effect. Analysis of intraday data highlights a period of 30 second intervals either side of news release as the occurrence of this phenomenon (Smales, 2012). Likewise, the release of scheduled earnings also portrays this behaviour with variables increasing to above average levels more than 15 minutes beforehand.

Sentiment Proxies

Research in this field began sometime before the technology was available to collect and score news sentiment automatically. Proxies were used instead of sentiment scoring systems, such as earnings announcements, company reports and court filings. The earliest work dates back to 1971 where Niederhoffer determined the relative importance of news and its effect on stock markets according to the font size of print used. Progression from these methods resulted in “information” as a substitute for news where investor sentiment was inferred, see Berry and Howe (1994) and Engle and Ng (1993). In fact, it is precisely these sets of “information” upon which investors’ determine their perceptions of the riskiness of an asset. Future prospects of a company can be deduced along with additional knowledge about the company.

Studies started to examine macroeconomic announcements by the 1990s with Ederington and Lee (1993, 1995) and Becker at al. (1996) leading the field. Cutler, Poterba and Summers (1989) introduce the concept of applying macroeconomic news to the variance of stock price returns. Surprisingly, their findings suggest that qualitative news stories unaccompanied by quantitative events do not explain large market returns. Bernanke and Kuttner (2005) and Rigobon and Sack (2004) are among the many papers that analyse the impact of changes in monetary policy and confirm a dynamic response in the markets to the release of (surprise) news. However, this effect quickly dissipates following news arrival. Later, work in this area advanced to employing news that was categorised into scheduled and non-scheduled news, or anticipated and non-anticipated news. Examples of scheduled news are earning announcements, dividend announcements and annual reports. Non-scheduled news is hard to work with because of its' noisy properties and the difficulties in quantification and interpretation. Lee (1992) and Li and Engle (1998) are examples of such work where the respective effects on stock returns and volatility are considered. Even earlier work on scheduled news dates back to Patell and Wolfson (1984) and Woodruff and Senchack (1988) where they notice that much of the market movement occurs in the first 30 minutes after corporate announcements.

Continuation of research on news stories introduced the categorisation of firm specific and non-firm specific news. In these instances only news detailing the events of a particular company is considered. Firm specific news has been found to drive movements in both stock prices and volatility as Tetlock, Saar-Tsechansky and Macskassy (2008) discover. Their quantitative measure of language produced forecasts for low firm earnings for a fraction of negative words. Nevertheless, reaction of stock prices to negative sentiment is brief and information retrieved from firm specific news stories are actually already priced in daily market prices. That is to say, no drift of prices occurs after the day of news release. Tetlock et al. (2008) argue that linguistic media content is advantageous in capturing aspects of firms' fundamentals that otherwise would be hard to achieve.

Preceding the quantified news sentiment data, researchers had to create innovative proxies to reflect sentiment in the markets, adopting measures from a range of asset classes. Firstly, stock option prices and trading volumes were seen as a source for extraction of sentiment. Whaley (2000) uses the market implied volatility index (VIX), also commonly known as the investor fear gauge, to test for risk aversion. Dennis and Mayhew (2002) utilise the put-call trading volume ratio as an index for sentiment in the derivatives markets. Their aim is to determine explanatory factors for the skewness in volatility of stock option prices with market sentiment as a variable to be tested. The motivation behind the inclusion of sentiment as a variable is that pessimism in the markets induces more skewness to the risk-neutral density, where increased volumes of put options indicate pessimism. However, results from cross-sectional regressions do not support such an explanation. Meanwhile, Kumar and Persaud (2002) employ the Risk Appetite Index (RAI) evaluated as Spearman's rank correlation of stock price volatility and excess returns. By purely monitoring the risk/return trade-off, a specific emphasis can be directed at the market's willingness to accept risk. Similarly Bandopadhyaya and Jones (2006) develop an Equity Market Sentiment Index (EMSI) composed of the Spearman rank correlation between the rank of daily returns and the rank of historic volatility with values ranging from -100 to 100. The advantage of such a measure is that changes to the underlying riskiness of the market do not directly affect the EMSI. Baker and Stein (2004) suggest liquidity as a type of sentiment index, in the form of share turnover, and argue that signs of high liquidity are a symptom of overvaluation. Baker and Wurgler (2006) also form a sentiment index to test their theoretical argument for irrational markets.

Alternatively approaches utilising debt-based securities and cash-flow of funds have also been explored. Randall, Suk and Tully (2003) measure net cash flow into mutual funds as a proxy for stock market sentiment. Lashgari (2000) uses the Barron’s yield spread ratio as well as TED spreads as indicators of market confidence. TED spread is the difference in interest rates on interbank loans and short-term U.S. government debt, subsequently forming the acronym from T-Bill and ED (the ticker symbol for Eurodollar futures contract). Narrow spreads signal high confidence in the markets and it is found that both the TED spread and the yield spread ratio can only explain 6% of variations in stock returns. However, the dominating issue with these sentiment indicators and those mentioned beforehand is that they are not orthogonal to asset prices, which is the variable that should be explained by the sentiment index.

By the turn of the century, sources of news information were flooding on to the World Wide Web and researcher were quick to jump on board this publicly accessible fountain of information. Das and Sisk (2003) utilize message posts and stock market discussion forums to mine for reasons explaining the impact of information on stock prices. Antweiler and Frank (2004) analyse internet stock message boards and quantify the language by categorizing the content as either “buy”, “sell” or “hold” recommendations. Evidence is found for an existing relationship between message activity, trading volume, return volatility and costs of trading.

News in Prediction Models

Tetlock (2007) was a pioneer in the exploration of news sentiment on stock market activity and was one of the first to directly use news sentiment as a factor. His source was a back page column of the Wall Street Journal in which he created a media pessimism factor. This measure is used as a proxy for market sentiment. The results presented have been heavily cited by fellow academics working in the field of news analytics because he claims to be able to predict stock market movements. After carrying out many regressions on vector autoregressive (VAR) models, he reaches the following conclusions. Firstly, there is evidence to justify the hypothesis that high media pessimism is associated with low investor sentiment resulting in downward pressure on prices. Statistically significant results indicate that high levels of media pessimism robustly predict downward movement of market prices. However, such negative influence is only short-lived and prices are almost fully reversed later in the trading week, consistent with the model of Campbell, Grossman and Wang (1993). This model supports the argument that sentiment theory proposes reversal of short-horizon returns in the long run, whereas information theory states that returns have indefinite persistence. Furthermore, the pessimism measure exerts an effect on returns that is an order of magnitude greater than typical bid-ask spread for Dow Jones stocks. Moreover, he finds evidence of longer-lasting and larger impact of negative sentiment on small stocks, measured according to the small-minus-big factor of the Fama-French model. Implicitly this can also be seen as a measure of individual investors' views. Additional findings include the rising of media pessimism when market returns are low and extreme values of pessimistic sentiment (high or low) leads to temporarily large market trading volume but this does not directly forecast volume.

A criticism of this work is the source used to construct the measure of media sentiment. The content of the newspaper column is in relation to the previous day's market activities meaning that the derived sentiment is based on information from past events. Thus investors will have already acted on such information and stock market prices are reflective of this. Tetlock argues this matter by performing robustness tests on the return window.

Barber and Odean (2008) investigate the trading decisions made once investors' attention has been caught by news, where the Dow Jones newsfeed is used as the dataset. A distinction is made between individual investors and professional investors where individual investors are more likely to purchase shares on high-attention days. Professional investors are less likely to be influenced by such events. Through observation of imbalances in buy and sell trades, Barber and Odean report a distinctly different behaviour of investors on stocks that appear in the news and those that do not. Additionally, imbalances are found to be greater on days that have negative return than positive return days implying greater trading activity during periods of decreasing stock prices.

The prediction of stock returns using news and its quantitative measures have not been extremely successful or reliable yet some have found promising results. Tetlock, Saar-Tsechansky and Macskassy (2008) apply the Bag-of-Words scheme to quantify sentiment of firm-specific news text and find that negative words in these news stories forecast low firm earnings and low stock returns on the next trading day. Furthermore, evidence is shown that integration of quantified language significantly improves forecasts of investors' reactions in stock markets, better than measures of market prices and analysts' forecasts. Sinha (2010) uses readily processed sentiment scores to construct a measure of long-term qualitative information to predict returns. His findings reveal an under reaction of the stock market to news and momentum is explained as the reason for this effect. Engelberg et al. (2012) investigate the negative relation between short sales and future returns. A significantly stronger link is observed in the presence of negative news. Dzielinski et al. (2011) found the relationship between news and return to be the following: positive news produces above average returns and negative news results in below average returns. Furthermore, comparing the average return of no-news stocks with the effect of neutral news it is established that the difference is non-distinguishable. On the other hand, Tetlock (2011) reports that investors do in fact react to stale news where staleness is defined as textual similarity to the previous ten stories about the same firm. This result is most prominent in individual investors who tend to overreact and trade aggressively on stale information causing temporary movements in stock prices. Such findings are contrary to stale/no information theory which predicts no effect of media pessimism on trading volume.

Most recently, Lee Smales from Curtin University, Australia has produced results relating news sentiment to trading activity and volatility movements, implementing the latest scoring system of sentiment. Smales (2012) identifies significant effects induced by contemporaneous news items for all variables (money value traded, volatility, absolute order imbalance, bid-ask spread and average trade size) apart from returns. Market activity, represented by money value traded and order imbalance, revealed a particularly significant and positive relationship with news while high-relevance news produce an increase in market activity as well as volatility and spread. A particular emphasis on the global financial crisis revealed an impact of news on market activity and volatility for four of the biggest banks in Australia (ANZ, CBA, NAB and Westpac) however this relationship was non-existent before the crisis.

Smales (2013) discovers that the relationship between news sentiment and VIX is asymmetric and significantly negative, that is, negative news events corresponds to large movements in VIX with a greater shift in prices induced by negative sentiment than positive sentiment. Furthermore, by examining the relationship annually it is determined that news sentiment has an increasing strength of relation with fluctuations in VIX when periods of implied volatility increase. VIX is used as a consensus towards expected future stock market volatility where large values represent greater fear. This index is estimated by averaging the weighted prices of puts and calls based on the S&P 500 over a wide range of strike prices. Moreover, experiments carried out in this study produced evidence suggesting that news is not endogenous. In other words, the lagged changes in news sentiment do not affect changes in VIX, particularly in intra-day intervals.

A common phenomenon detected in the latest research on news sentiment analysis is the reaction of prices occurring before the release of news (see Figure 1.1). Explanations to this phenomenon include the presence of informed traders or the clustering of news arrivals, with the former deeming more likely. Evidence found by Gross-Klaussman and Hautsch (2011) and Smales (2012) show that market makers do not adjust their spreads during these periods prior to news release, which according to Kyle (1985) reflects the presence of informed traders. Moreover, average trade size does not increase significantly either, suggested by Easley and O'Hara (1987) as another indicator of private information. An alternative possibility could be information leakage before the news announcement, either by a competitor within the industry or an individual who has obtained the information. Interestingly, this highlights the requirement to study social media where many influential individuals congregate and share what is known at that stage as rumours. The investigation into such a relationship could resolve this pre-news reaction in markets.

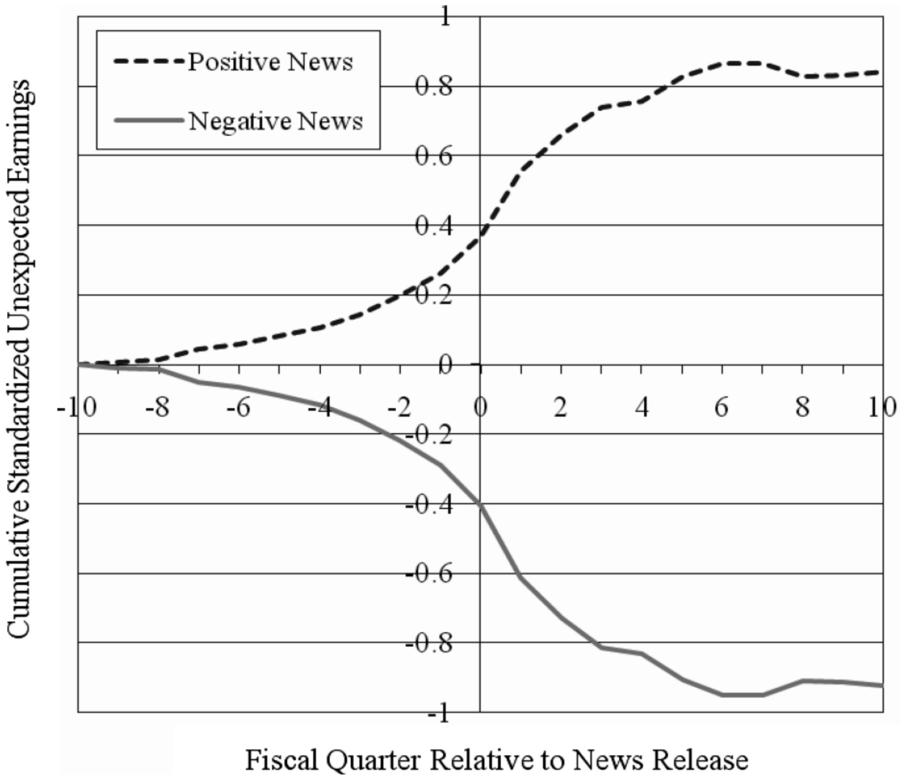


Figure 1.1: Firms' cumulative standardized unexpected earnings 10 fiscal quarters before media coverage of an earnings announcement until 10 quarters after the media coverage. The figure portrays the change for both positive and negative news stories. Source: Tetlock, Saar-Tsechansky and Macskassy (2008).

1.4 The Power of Unstructured Text

Sources of Information

With the development of different communication mediums, caution needs to be applied to the sources used to derive news sentiment. Only an accurate reflection of news content may lead to reliable and convincing research results. Using trustworthy sources such as Dow Jones newswire, Thomson Reuters' newsfeed, Bloomberg newsfeed and the Wall Street Journal guarantees reliability of official information, whereas discussion forums and online blogs may be more appropriate for aggregation of sentiment in the general public. Modern technology has also introduced the world of social media; micro-blogs and Twitter are gaining popularity. To be precise, Twitter is in fact a micro-blogging platform. Together with newswires, these sources of news are considered to be a "push" of information in the sense that such sources aim to spread and publicise the content within. As these online communities expand, the information shared is also known to impact markets. One of the earliest publications relating Twitter data to stock markets is Zhang, Fuehres and Gloor (2011). Simply collecting data for six months, emotion measures of hope and fear were found to be correlated with American indices Dow Jones, NASDAQ and S&P 500. Bollen, Mao and Zeng (2011) research the predictive implications of large scale Twitter feeds on the closing values of the Dow Jones Industrial Average (DJIA). Through tracking the mood of daily Twitter feeds they find 86.7% accuracy in predicting the daily up and down movement of closing values, although only specific mood dimensions achieve this result. The sentiment derived from social media platforms such as Twitter is subtly different from news sentiment. An alternative term would be public mood, as the messages shared are more personal and reflective than reported news.

Additionally, news sources that “pull” information also exist such as websites that scrape for online information. Rather than spreading first-hand information, these sources gather the information, analyse and investigate it before providing summaries of the findings. An example is *Google Trend* where analysis of voluminous search queries present information about concerns and interests of the general public. Research conducted by Preis, Moat and Stanley (2013) received outstanding media attention recently for their revealing findings that warning signs of market movements can be detected using *Google Trends*. They argue that a combination of extensive behavioural data and market data provides a better understanding of collective human behaviour.

Since the validity of information from such sources cannot be guaranteed, there is growing research to process and filter such information stream and enhance the trustworthiness of the contents. But this line of investigation is outside the scope of this thesis.

Motivations for conversion of machine-readable news to sentiment scores

Theoretically, the value of a firm is equal to the expected discounted value of cash flows conditional on investors’ information sets and it is the source of these information sets that we are interested in. The analysis of quantitative information is very well established; in contrast the exploration of qualitative information is still at an early stage. However, there exist several compelling reasons to investigate the unstructured text. First, unravelling the content of news brings an unlimited variety of events that can be examined for impacts on the stock market, without any restrictions like those of previous research (Ederington and Lee, 1993, Becker at al., 1996). Second, investors' impressions on firms' fundamental values are mainly judged on second-hand information and not through direct observation of production activities – one of these sources is the media. Aspects of firms' fundamentals captured in news text are otherwise hard to quantify through alternative sources such as analyst forecasts and accounting variables. Therefore, to obtain an accurate valuation, including earnings and stock returns, detailed studies need to be carried out on the messages transmitted through the communication of the media – more specifically the news.

As stated earlier, news stories, unexpected or anticipated, create opinions and sentiment among investors which in turn lead to actions taken on trades. To capture this interpretation of text to opinion the field of sentiment analysis is introduced. Also known as opinion mining, the subjective content of text materials is identified and extracted using techniques such as natural language processing, text analysis and computational linguistics. News analytics utilises many areas of mathematics and computer science to summarize and classify public sources of information for instance machine learning, collaborative filtering, information retrieval and statistical learning theories. Previously researchers have had to use proxies for sentiment because simplified quantitative data was not available but very much in demand. With demand comes supply and many companies are now in the business of providing news analytics data. The process of collection, tagging, aggregation through to scoring has now been fully automated so that live feeds of sentiment scores along with the news can be provided. Language processing algorithms are set up to handle intricacies of the human language within written text. Armed with this new dataset of sentiment scores, researchers are now equipped with a wealth of knowledge regarding the behaviour and beliefs of investors. If this score varies significantly over time, then market beliefs about the company are also changing quickly.

The Content

News items may not solely talk about a single company or be the latest article discussing a certain issue therefore these features need to be sorted and ranked, quantitatively if possible. Many data fields arise when categorising news, for example importance, novelty, relevance and credibility. There are plenty more descriptive fields. In Chapter 3 we explain all of these in more detail. Filtering by these characteristics can reveal trading signals or potential risk warnings. In order to affect stock returns, the piece of information conveyed needs to be novel.

Not only can the interpretation of words used in news articles be meaningful but also the tone of the author; essentially the author’s sentiment. This is reflected in the choice of words used and is the starting point to sentiment analysis.

Sentiment analysis does not only look at the text in news articles but also considers the entire content and hence derives the context in which they are applied. The classification of positive and negative sentiment in text is a common two-class problem in sentiment analysis (Pang, Lee, et al., 2002, Turney, 2002).

Conversion to quantitative measure for analysis

Computer scientists have been able to transform news text into numerical values for analysis using linguistic pattern recognition tools. This conversion of qualitative data to quantitative data opens a whole new world of research allowing news sentiment to be a direct input into mathematical models. Newsfeeds nowadays appear on traders screens within a matter of milliseconds. This speed of messaging is easily linked to automated trading and high frequency trading. With all scoring systems set in place electronically, the conversion rate of machine readable news to a sentiment score is in line with such trading frequencies. Furthermore, it is reasonable to link news analytics to this form of trading, as many studies have shown, since the impact of news on asset prices is already incorporated at lower frequencies such as daily trading. Hence, naturally the obvious route is to study it at an intra-day frequency.

As described in section 1.3, there are several definitions to the term sentiment and with this evolution brings change in the classification process of sentiment. Initial classification methods merely counted the number of words related to positivity and the words related to negativity within a short piece of text, limited to approximately 50 words. Examples of studies that applied this process are Li (2006), Davis, Piger and Sedor (2006) and Tetlock (2007). At this stage only two types of sentiment were considered; neutral sentiment did not exist. Categorisation of positive and negative words was decided through predetermined databases that had already assigned polarity to a large selection of words, e.g. the General Inquirer (GI) of the Harvard Dictionary. The GI is a well-known quantitative content analysis program designed by psychologists spanning 77 categories in total. Tetlock (2007) extracted media sentiment using this process and evades text mining by converting columns of text to numerical values and undertook principal component analysis. Similarly, Tetlock, Saar-Tsechansky and Macskassy (2008) produce document-term matrices filled with frequencies of word appearance relative to a full piece of text; a common scheme known as Bag-of-Words. Focusing on negative sentiment, they selected all words falling into the predetermined negative categories of the GI and considered them equally informative, and subsequently summed everything to determine the degree of negativity in a news article. However, these scoring systems cannot be completely relied upon. Loughran and McDonald (2011) found that three-quarters of words identified as negative in the Harvard Dictionary are not typically considered negative in a financial context. Therefore, there remains a requirement for the presence of human judgement alongside the large databases capable of distinguishing overall sentiment of news stories.

However, dictionaries and databases that only assign sentiment to singular words do not fully interpret the content in news articles. Complete sentences and phrasing construction need to be considered in order to discover the exact context in which words are used. Only then can a more accurate reflection of the news sentiment be deduced. Hence, grammar practices such as negation and adverbs are introduced to the sentiment classification process so that scoring is applied to a string of words and then an average score is taken as the overall sentiment. Now, it is common practice to categorise this step as part of text pre-processing, with Das and Chen (2007) introducing the first negation tagging method.

The techniques mentioned so far are language dependent i.e. domain knowledge is required to carry out tasks. Contrasting to these methods is a group of classifiers that do not require any predefined knowledge e.g. Bayes Classifier and Support Vector Machines. The most widely used classifier in practice is the Bayes classifier, which has many different versions. It uses word-based probabilities and pre-classified text to assign a category to new text. Specifically, a corpus of news is initially accurately classified to be used as training data to identify the prior probabilities, which form the basis for Bayesian analysis. Posterior probabilities of categories are determined by applying the classifier to out-of-sample data, with the assignment of the specific category decided by the highest probability. Alternatively, discriminant-based classifiers have also been introduced, which adjust term weightings to identify the more emotive words.

Similar to these methods and also taken from the field of computer science are machine learning and natural language processing, where a training set of data is required to initiate the algorithm. The structure of these algorithms involves creating a set of already classified news text by humans to form the training set, on which the computer detects patterns and records them. Using this information it is then able to classify the news through learning and updating these patterns. Finally metrics are applied to assess the accuracy and stability of results. Additionally, machine learning algorithms may also be applied to identify relevant tags for a story. These tags turn the unstructured stories into a basic machine readable form. The tags are often stored in XML format and reveal the story's topic areas as well as other important properties. For example, they may include information about which company a story is describing. The basic idea behind these technologies is to automate human thinking and reasoning.

Automation of the classification process provides many benefits. For example, excluding humans in the judgement process generates a higher degree of consistency in results because no emotion is involved. Pang, Lee and Vaithyanathan (2002) compare classification results for humans and machine learning techniques and found that indeed machine learning performs better. However, simply applying a singular technique such as naïve Bayes or support vector machines did not beat the performance of topic-based categorization. Hence, modern day classification techniques adopt a combination of new methods, such as Bayes classifier, and existing ones, such as text mining.

Sentiment Classifiers

All classifiers can be separated into two broad categories: supervised and unsupervised learning methods. Classification algorithms that are based on well-defined input variables belong to supervised learning, whereas unsupervised methods refer to those latent variables still to be found. The former group of methods are well understood and deeply researched but the latter are less explored in comparison, although its popularity is increasing. Examples of unsupervised learning techniques are cluster analysis and community detection, which is where the focus of news analytics is converting.

Nowadays, news sentiment data sold by major data vendors such as Thomson Reuters and RavenPack apply a combination of tagging methodologies to achieve the most accurate reflection of news content. Traditional tagging algorithms plus expert consensus and market response methodologies guarantee a high degree of accuracy. The addition of manual tagging by humans, although tedious, improves results significantly as it constructs the foundations upon which machine learning processes are built. Training from a set of true reflections in sentiment minimizes errors in prediction. In fact, such a process was adopted from the PR/Marketing industry that used media tracking to compute reputations of companies.

Three primary steps are involved in the transformation of qualitative text to quantitative scores. They are:

(i) Tagging process
(ii) Sentiment classifiers
(iii) Score calculation

The initial tagging process is very important in distinguishing the key attributes of the news stories and subsequently the classification of its sentiment. Some of these aspects include the entities to which the story is relevant, topics covered, and the market it applies to.

Next, using story type as a preliminary step, initial judgement of sentiment polarity is made. RavenPack utilises two main methods in detecting sentiment – the Expert Consensus Method and the Traditional Method. The former is ultimately employed to train Bayes classifiers to imitate the tagging rules of experts by training them on several thousand news stories that have been manually tagged. The Traditional Method maps words, phrases, combinations and other word-level definitions to pre-defined sentiment values. This technique is an advancement from the consideration of only singular words. Both algorithms are composed of several steps with the opening task of defining a Classification Base, followed by the construction of a Rule Base or Tagging Guide, which is then tested on a large sample for accuracy. The data series is only ready for publication after consistency checks of historical data and generation of volume statistics. To maintain the most up-to-date language patterns in the tagging methodologies, RavenPack carry out re-evaluations every quarter.

Finally, the sentiment score of specific companies can be generated through aggregation of the individually tagged stories based on various sentiment classifiers. Thus far, all sentiment has been specific to a news item and not been separated at the entity level. Hence by accounting for the relevant companies and sectors, a score is calculated as a weighted average of a selection of sentiment classifiers. Relying on purely one classifier can be too erroneous. Some classifiers may perform better in topic-based categorisation whereas others might be stronger in sentiment classification; hence an amalgamation achieves optimal results and improves the signal-to-noise ratio.

Measures of news impact on markets

Easley et al. (1996) derived a measure known as Probability of Informed Trading (PIN) to investigate the impact of information-based trading on liquidity spreads. A key empirical result is that the probability of information-based trading is lower for high volume stocks, tested on a selection of stocks that are commonly and uncommonly traded on the NYSE for the year 1990. Coincidentally, the PIN metric is also an indicator of how much information has been digested by the market through changes in the bid-ask spread. If the spread differs in size to the period before information release, then it can be assumed that traders have taken in and acted on such news. In general, a narrowing of the spread denotes upwards price movement and therefore good news, whereas widening of the spread represents a drop in price and hence bad news. To determine the effect of information on spreads only four parameters are required: (i) the probability of new information occurring, (ii) the probability of bad news appearing, (iii) the arrival rate of uninformed traders according to a Poisson process, (iv) and the arrival rate of informed traders also according to a Poisson process. The analysis is performed on groups of stocks according to their trading volume with differing results between stocks that are actively traded and those that are inactive. Furthermore, using regressions, a relation is established between PIN and spreads, that is, the greater the probability of informed trade, the larger the spread. Fluctuating bid-ask spreads can also be explained as the market maker’s perception of risk on information-based trading. Therefore, a larger spread may be priced to act against informed traders for example. Unlike Kyle (1985), Easley et al. (1996) do not aggregate buy and sell trades therefore it is only the composition and number of trades that determine beliefs.

1.5 News and Its Use in Fund Management and Trading

Models that effectively incorporate news data for decision making in trading strategies and fund management are growing in interest. Fund managers are required to select portfolios in anticipation of profitable returns. Ideally they would like to foresee the price movement of each asset in a portfolio for some period in the future, but unfortunately for us mere humans, the best we can achieve is through prediction. By overcoming these uncertainties about asset behaviour, only then can one make reasonable judgements on trading choices. To date fund managers have market insight, accumulated over years of experience, and information gathered through public sources neither of which are solid enough to make decisions with 100% confidence. Applying a fusion of market data and news data to predictive models enhances predictability through the addition of content value. Tetlock, Saar-Tsechansky and Macskassy (2008) develop a fundamental factor model that incorporates news as a factor. Excluding consideration of transaction costs, they build a profitable trading strategy based on this model.

An alternative application of news in the financial markets is for the surveillance of trading activity. Circuit breakers may be implemented according to live feeds of news data so that once a certain threshold is reached a trigger alerts traders of extreme contextual polarity, whether it's in the sentiment score or novelty score or another factor. This could lead to human interference/supervision in algorithms to cancel or change orders due to new information which alters positions in the markets. For example, negative sentiment from a news item can protect the trader from being blindsided by an event if an aggressive purchase was planned, or conversely, increasing a purchase in the presence of a highly positive sentiment may reduce slippage and lower transaction costs. It is also possible to incorporate news sentiment into trading strategies, either to improve performance or find alternative trading signals. Schumaker et al. (2012) identify possible contrarian trading behaviour, with their system able to predict price decreases using articles with positive sentiment and price increases through negative sentiment. Furthermore, they found that subjective news articles and negative sentiment are more successful in the prediction of price direction.

The financial markets are converging towards a fully electronic trading place whereby all transactions of asset classes such as equities and foreign exchange are processed through electronic platforms. Such vital conversions have ushered traders to rally behind a form of trading known as automated trading, which relies heavily on computers from decision making to processing order executions. Since automated trading is growing in its adoption of news sentiment, we discuss further the relationship between them in Chapter 2.

Additionally, systems can be implemented to exploit the volatility surrounding significant news items. Besides the applications of news sentiment mentioned above, risk managers have also seen potential in news sentiment to improve their models. Academic research has repeatedly proven that incorporating sentiment to volatility predictions significantly increase accuracy (Mitra, Mitra and diBartolomeo, 2009, Smales, 2013). Improving risk estimates is an outstanding objective for risk management departments. It is commonly known that return predictions are hard to model accurately and this is also the case when applying sentiment for prediction. However, other means of signals may be found in volatility or liquidity measures which are known to have better performance in prediction models (Gross-Klaussman and Hautsch, 2011, Smales, 2012). Being able to anticipate any unforeseen risk is an advantage whether it is directly from news of the company or its peers.

Not only can news sentiment be considered as a variable in predictive modelling, but many other practical applications exist as well, for example, trading as part of an index.

Motivations

This thesis investigates the application and exploration of news metadata in high frequency, precisely one minute bar, in the field of news sentiment. The motivation for such a choice is justified in detail in Chapter 2 and based on an established result of the inevitable reversal of stock price returns after news release. Some have found this period to be a few days after the news event (Tetlock et al., 2008) whilst others say it is longer – around 10 days (Seasholes and Wu, 2004) to a month (Uhl, 2011).

However, few have explored the alternative end of the spectrum and refined analysis to a magnified intraday period surrounding news release, i.e. minutes before and after. Furthermore, a spotlight has been cast on a particular form of trading in these timeframes, namely, automated trading. Several events occurring on the stock markets have highlighted the role of these traders and brought it to the tip of everyone's tongue as a hot discussion topic. One crucial event was the "flash crash" of 6 May 2010, where the Dow Jones Industrial Average plummeted 1000 points in the space of minutes. Shockingly of all was the fact that most of the losses were recovered within minutes. Strong criticism fell on automated traders for their participation in the events that day and their style of aggressive trading led to the fall in stock prices. Cases such as this highlight the need for detailed research in automated trading to better understand the behaviour of markets and assets under these conditions.

Researchers have studied the relationship between news sentiment with stock price returns and volatility, yet minimal attention has been put on liquidity. However, in this thesis we explore all three measures simultaneously and observe the respective effects of sentiment on each measure. The reason why we include liquidity in our work is instigated by the consequences of the global financial crisis in 2008-2009. With the collapse of Lehman Brothers in 2008 and the subsequent banking crisis and bank bail outs in the UK and other European countries, the importance of liquidity has become paramount and is widely acknowledged by the finance community. Implications of the financial crisis include regulators imposing stricter liquidity requirements for banks, especially those active in low latency trading activities (Gomber et al. 2011, AFM 2010), and triggers and restrictions being implemented in stock markets as a prevention for future crashes. Consequently with regards to trading activity, the availability of liquidity to traders and brokers have become of utmost importance. In this thesis we consider (i) news and its impact with liquidity and (ii) relevant liquidity measures (see Section 2.3).

In order to understand liquidity well, the theories of market microstructure need to be introduced and explained. The study of market microstructure started around four decades ago and has attracted further attention in the past decade with the advent of computer-driven trading and the availability of all trade and quote data in electronic form, leading to a new field of research called high frequency finance. Research in high frequency finance demonstrates that properties derived from low frequency data to define the behaviour of financial markets fail to explain the market behaviour observed in high frequency. Three events are cited as early triggers for the general interest in microstructure (Francioni et al, 2008):

1. the U.S. Securities and Exchange Commission's Institutional Investor Report in 1971;
2. the passage by the U.S. Congress of the Securities Acts Amendment of 1975; and
3. the stock market crash in 1987

1.6 Thesis Outline and Contributions

To date, prediction and modelling of stock market prices have primarily adopted market data as the dominant source of information retrieval, in the form of open, high, low and close prices, bid and ask prices and traded volume. A contribution of this thesis to the field of predictive analytics is the fusion of the content in news events into the prediction process, more concisely, news sentiment metadata. Intelligence garnered from newly available news metadata permits the incorporation of investors' sentiment and reaction to non-anticipated events and announcements. Moreover, the studies presented are set in a high frequency timeframe, that is, minute bar frequency to provide another innovative feature of our work. Currently, research lacks literature that examines the effects of news sentiment on stock price returns, volatility and liquidity in such a magnified frequency. Additionally, we emphasize the increasing importance of liquidity whilst considering trading activities by including it as an asset behaviour variable to be modelled.

In Chapter 1 we have presented the focus of the thesis in a summary form and introduced the concept of news as an event. We also described how the study of (news) sentiment has evolved, the essential aspects of analysis and classification of unstructured text. Finally the use of news and its applications in fund management and trading are discussed.

The rest of the thesis is organised in the following way.

Chapter 2 describes automated trading, market microstructure and liquidity. Given the speed of arrival of news feeds as streaming data and the computer processing ability, it is easily seen that automated trading and streaming news are two complementary processes. Supplementary information provided by news can be incorporated into established strategies such as fundamental and technical analysis to improve trading performance. In order to understand the intricacies of trading mechanisms and the behaviour of market participants, a brief summary of market microstructure is given in section 2.2. By exploring the field of market microstructure, the requirements of liquidity by market participants are highlighted. An explanation of liquidity and its proxy measures are given in section 2.3.

In Chapter 3 we first present the news metadata in a summary form and also provide alternative ways of assigning sentiment scores. In particular, we introduce the concept of “impact of news”. The related impact score is defined such that it takes into consideration the decay of sentiment as well as the accumulation of sentiment for multiple news stories. The development of the Impact score was instigated through the exploration of news volume, better known in this field as news flow. This measure is novel and is one of the major contributions of the thesis.

In Chapter 4 we propose univariate prediction models for stock return, volatility and liquidity, as representations of asset behaviour, using the novel measure of the Impact score. The study is based on equities because the majority of news is directed at individual companies. The time frequency of our selected data is minute-bar and hence can be considered as high frequency. This feature along with the prediction of liquidity as a trading measure provides novel contributions to the field of predictive analytics as these are both uncommon territories lacking in research. Autoregressive (AR) models and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models are adopted to predict the independent variables (asset behaviour measures). Empirical results show improvement in prediction performance of liquidity and volatility by applying news sentiment, in the form of the impact score.

Experiments of multivariate models are carried out and are explained in Chapter 5. Employing the approach of the Black-Litterman model (1991), we consider news sentiment as a prior belief and build the posterior distribution based on this. Bayesian inference techniques were adopted to better incorporate the effects of news sentiment on asset behaviour. The results obtained are analogous to those in the univariate setting but with less prominence.

Chapter 6 concludes and presents ideas for future research.

Chapter 2

Market Microstructure, Liquidity and Automated Trading

2.1 News and Its Relationship with Trading

Given the impact of news and news flow on asset prices (return and volatility), automated traders wish to incorporate machine-readable news into their trading strategies automatically. This would imply that thousands of sources of information can be simultaneously screened and their conclusions derived within a matter of minutes, freeing up time for the trader to handle more complicated tasks that would require the human brain. Another advantage of automating processes is their ability to follow instructions precisely, efficiently and emotionlessly. Much of trading is now occurring electronically with equity and foreign exchange markets all transacting on electronic platforms. Thus it is only logical to combine electronic news feeds with automated strategies to maximize the content of information and position themselves correctly in the markets. The procedure of news dissemination, collection and quantification are all fully automated, complying with the data processing speed of automated trading. Thus, it is a natural choice to combine the two procedures to attempt to achieve improved trading performance.

Speed is a major benefit for automated traders, allowing them faster access to information, decisions made quickly and rapid position taking in the markets. According to market microstructure theory, investors have always lost to better informed traders (O’Hara, 1995). However, for the case of automated traders, their advantage is not that they are better informed but simply their speed in receiving the same information. This is a fact that has angered retail traders. Proposals have been suggested for methods to protect investors from such losses e.g. notification of exchanges to halt trading of equities for those companies who are about to release news. Similarly, many governments already follow this rule and only reveal major news when markets are closed or at pre-announced times (Harris, 2013). Furthermore, studies have found a mispricing of assets surrounding news release periods thus presenting opportunities for arbitrage. Gidofalvi and Elkan (2001) identify this period to be around 20 minutes. If this is a true reflection of markets then there certainly exists an urgency to act on news. The efficiency of automated trading satisfies this condition. Trading activity by automated traders following news releases allow prices to be made more efficient by the second. However, this does not offer any economic benefits; the effects are on a microstructure level as discussed in more detail in section 2.2.

To facilitate the fundamental pricing of securities, computer systems are set up to collect news announcements and similar forms of information. With the motive of enhancing price movement forecasts, low latency trading firms integrate this information provided in news content and trade accordingly at the quickest moment. Harris (2013) identifies a trading strategy of automated traders that trades on fundamental values derived from news feeds. He explains that some traders install computers to monitor and interpret electronic news feeds, which in turn are able to identify relevant information at lightening speeds facilitating immediate determination of optimal trades. As a consequence, asset prices are likely to reflect fundamental values information faster than it otherwise would.

Information portrayed through news prompts two different types of trades - those who “wish to trade” and those who “need to trade”. Broker-dealers are an example of traders who after receiving information decide that taking a trading position on such information would be beneficial for them. Hence, they possess a wish to trade triggered by the information obtained and ultimately seek excess returns (alpha). On the other hand, market makers are an example of traders who need to fulfil market/limit orders and prevent drastic changes in asset price, therefore having an obligation to trade. In these instances, the optimum action such traders can take is to execute at the best price for their positions.

Moreover, a Foresight Project conducted by the UK Government Office for Science titled “The Future of Computer Trading in Financial Markets” also further highlights the concern of such issues by the government and regulators.

Armed with massive databases of historical data, a selection of traders nowadays are able to produce quantitative models unavailable to other parties because of this rich resource. Programmed so that they can run algorithmically, these quantitative models produce fast, efficient and emotionless decisions making it a requirement in the operations of the modern day speedy markets. Nevertheless this does not imply that traditional methods such as technical analysis and fundamental analysis are completely eliminated from trading models. LeBaron, Arthur and Palmer (1999) find from their artificial stock market of simulated traders that those who react quickly to new information introduced to the markets mostly apply technical strategies, whereas those with longer waiting times formed fundamental strategies.

Not all asset classes are tradable using automated strategies so firms need to select markets that satisfy certain conditions, mainly associated with the market itself. One of these conditions is a highly liquid market so that traders are able to quickly enter and exit positions, which is a crucial criterion underpinning the strategies of automated trading. More specifically, Black (1971) pointed out the presence of several necessary conditions for a stock market to be liquid:

(a) there are always bid and ask prices for the investor who wants to buy or sell small amounts of stock immediately;

(b) the difference between the bid and ask prices (the spread) is always small;

(c) an investor who is buying or selling a large amount of stock, in the absence of special information, can expect to do so over a long period of time, at a price not very different, on average, from the current market price; and

(d) an investor can buy or sell a large block of stock immediately, but at a premium or discount that depends on the size of the block – the larger the block, the larger the premium or discount.

Other properties include sufficient market volatility to ensure that changes in price exceed transaction costs thereby making it possible to earn profits, and an electronic market to enable the quick turnover of capital and to harness the speed of trading. Currently, only spot foreign exchange, equities, options and futures markets fulfil such conditions.

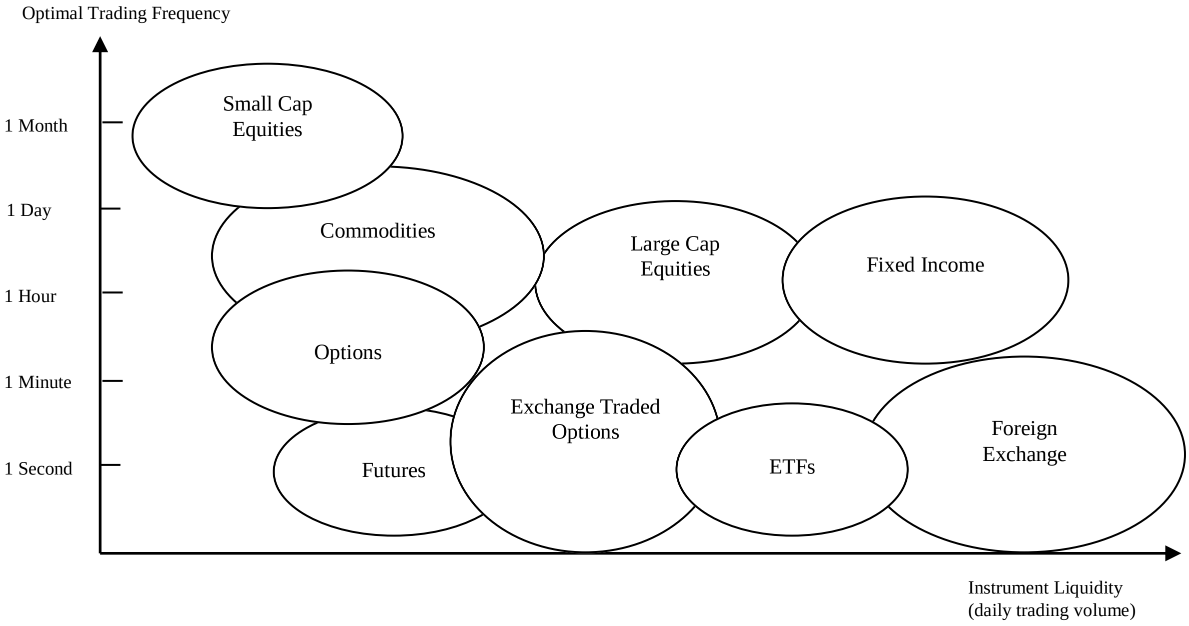


Figure 2.1: The trade-off between optimal trading frequency and liquidity for various trading instruments.

Figure 2.1 illustrates the trade-off between trading frequency and liquidity for many asset classes, using daily trading volume as a proxy for liquidity. Asset classes closer to the origin of the graph are more ideal for automated trading. That is to say, these asset classes which trade at an optimal frequency of less than a day tend to be accompanied with higher levels of liquidity, satisfying market requirements. This becomes an optimal trade-off for investors as it allows them to trade without having to worry about transaction costs diminishing their profits. This explanation is justified by how liquidity is defined: “the ability to convert assets into cash at the lowest possible transaction costs”. Furthermore, it can be noted that highly liquid asset classes can also be executed electronically and traded on a regular basis. In the markets today, we acknowledge that a sweeping but steady transition has occurred with the conversion of over-the-counter markets into electronic markets to keep up with the trading strategies of investors. We discuss in more detail the three major asset classes traded using automated strategies.

Equity markets

Equities are the most favoured asset class because of the market's large size and volume; this is supported by the market's breadth of listed stocks and its ability to be traded electronically. Moreover, such a breadth often presents market inefficiencies which are great trading opportunities. It is also popular for its diversification properties in portfolio investment with possibilities to hold long or short positions in stocks. In addition to stocks, the equity market also trades exchange-traded funds (ETFs), warrants, certificates and structured products. In particular, hedge funds are especially active in trading index futures. According to research conducted by Aite Group, the asset class that is executed the most algorithmically is equities; for instance, by 2010 an estimated 50% or more of total volume of equities traded were handled by algorithms.

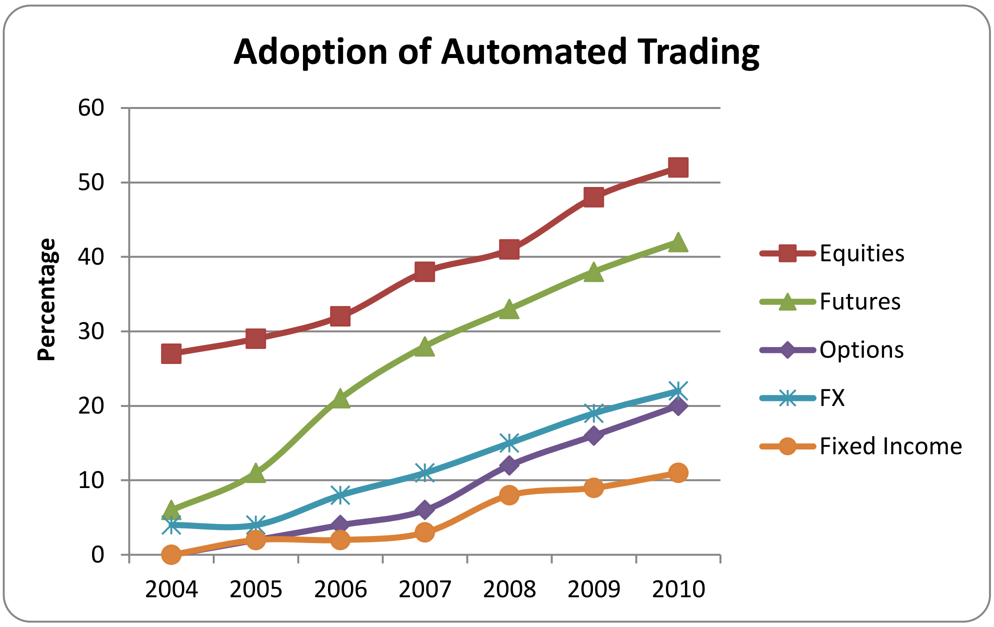


Figure 2.2: Progress in adoption of algorithmic execution by asset class from 2004 - 2010. Source: Aite Group.

Foreign exchange markets

The foreign exchange markets lack volume measures and the rule of "one price" because of its decentralised and unregulated mechanism. This has beneficial implications for automated traders as there are substantial arbitrage opportunities that can be identified by their automated strategies. However, there are only a limited number of contracts that can be found on the exchange, restricting the variety of financial instruments available for traders in the foreign exchange market. They are namely foreign exchange futures and select options contracts. Yet these instruments are traded electronically permitting them to be traded at frequencies under a minute.

According to Chaboud et al. (2013), algorithmic traders in the foreign exchange market monitor macroeconomic news meticulously so that a reduction in supply of liquidity can immediately be transacted after economic news arrivals to avoid being adversely selected. They conclude that following information shocks more liquidity is provided by high frequency traders than human traders. Furthermore, they find that trading at low latencies does not cause an increase to volatility in the markets.

Fixed income markets

Traders can profit from the fixed income markets by building strategies to take advantage of short-term price deviations in the pre-specified amounts paid to their asset holders. This asset class includes the interest rate market and the bond market but transactions are still taking place over-the-counter (OTC) indicating that trading frequencies normally occur daily although intraday trades can take place. The bond market contains an advantageous breadth of products with bond futures contracts standardised by the exchange and are often electronic. The most liquid bond futures are those which are nearing their expiry dates compared to bonds with longer maturities. The most liquid futures contract in the interest rate market is short-term interest rate futures with the bid-ask spread being the most common liquidity measure.

2.1.1 Trading Approaches Influenced by News

Handing over the role of trading execution to computers removes the element of human error stemming from emotion and bias. Downsides of such automation are the lack of intuition and ability to recognise unfavourable events where trading should be halted. As a matter of fact, with the development of text analytics and sentiment analysis computers are now able to incorporate news events into their systems to affect trading decisions. The conversion of text to numeric data in the form of sentiment scores aids the processing and consideration of news into trading algorithms, at a pace faster than humans can process the information. In essence, there should be a parity of emotion and unbiased judgement.

Hafez and Xie (2012) create a simple sentiment based strategy and execute it for the sample period of 2000 – 2011. They find that the strategy generates the greatest annualised returns during periods when the markets are bearish, i.e. August 2007 – February 2009. This behaviour is explained by them as the conveyance of negative news exposing more credible signals.

As previously mentioned, automated trading is not applicable to all asset classes as there are a set of criterion that needs to be fulfilled in order to implement a profitable strategy. Hence, there are trading strategies developed to manipulate these conditions and gain excess returns.

The trading strategies employed by automated traders can be distinguished as two groups: passive strategies and aggressive strategies. They can be defined by the types of orders utilized. Passive strategies exclusively place limit orders whereas aggressive strategies use market orders only (Kearns, Kulesza and Nevmyvaka, 2010). The difference in strategy implementation also affects the profiting technique, with aggressive traders carrying the burden of transaction costs and passive traders acting the role of liquidity provider thus taking on more risk.

Often, simple arbitrage algorithms are adopted which can also be performed at low frequencies. It is the speed of execution that offers the competitive advantage. Furthermore, trading approaches can be categorised into the following five types:

(i) Algorithmic Executions
(ii) Statistical Arbitrage
(iii) Event Arbitrage
(iv) Electronic Liquidity Provision
(v) Predatory Trading

Algorithmic executions refer to the process whereby a large order of stocks is broken into many smaller orders to be executed separately over several hours or even several days. This type of trading exists because of two reasons. Firstly, it would be almost impossible to find a counterparty that was willing to trade the exact same large order. Secondly, the breakdown of large orders into smaller chunks will significantly decrease the impact on the market, in terms of asset price and market conditions. For example, if a trader wishes to buy five million shares of a particular equity, the optimal strategy determined by the algorithm might be to purchase the shares over three trading days, breaking the large order into many small orders (i.e. 200 shares on average). An alternative way of executing this would be a principle bid trade with an investment bank, but such liquidity provision often comes at a high price. Not only do the analytical algorithms ascertain the size of orders but also the timing between small orders to reflect changes in the asset price, general market conditions, or the underlying investment strategy. Note that algorithmic executions only decide how best to execute an order and does not govern the nature of the trade (total number of shares or date and timing). Moreover, this class of algorithmic execution can benefit from the inclusion of news analytics and predictive analysis of liquidity.

Unlike algorithmic executions, statistical arbitrage trading automates the whole investment decision process and calculates deviations from equilibrium. A simple example of statistical arbitrage is “pairs trading”. Let us assume we identify the relationship that “Shares of stock X trade at twice the price of shares of stock Z, plus or minus ten percent”. If the price relation between X and Z goes outside the ten percent band, we would automatically buy one security and short sell the other accordingly. If we expand the set of assets that are eligible for trading to dozens or hundreds, and simultaneously increase the complexity of the decision rules, and update our metrics of market conditions on a real time basis, we have a statistical arbitrage strategy of the modern day. The most obvious next step in improving our hypothetical pairs trade would be to insert a step in the process that automatically checks for news reports. The system would be alerted when an indication of possible change in the monitored price relationship might occur as a result of a clear fundamental cause, as opposed to random price movements such that we would expect the price relationship to revert to historic norms. Pairs trading may also benefit by taking into consideration market sentiment as determined by news.

Event arbitrage is an automated trading strategy that manipulates the reaction of the stock markets to information events, i.e. macroeconomic announcement or news release. Event arbitrage strategies follow a three-stage development process:

(i) identification of the dates and times of past events in historical data
(ii) computation of historical price changes at desired frequencies pertaining to securities of interest and the events identified in the step above

(iii) estimation of expected price responses based on historical price behaviour surrounding the past events.

Therefore, by observing recurring events high frequency numerologists are able to take advantage of predictable short-term responses in the markets and generate short-term profits.

Electronic liquidity provision is a direct descendent of traditional over-the-counter market making, where a financial entity has no particular views on which securities are overpriced or under-priced. This renewed competition amongst liquidity providers, which previously was only supplied by specialist firms, means reduced effective market spreads (see Section 2.3) and hence reduced indirect market costs for final investors. Quantitative algorithms are programmed by high frequency firms to optimally price securities and execute market making positions. In order to do so, a thorough understanding and precise modelling of the target market microstructure is required. Typical holding periods only last one minute or less. The trader (liquidity provider) is automatically willing to buy or sell any security within its eligible universe at some spread away from the current market price upon counterparty request. Electronic liquidity providers differ from traditional market makers in that they often do not openly identify the set of assets in which they will trade. In addition, they will often place limit orders away from the market price for many thousands of securities simultaneously, and engage in millions of small transactions per trading day. Under the regulatory schemes of most countries such liquidity providers are treated as normal market participants, and hence are not subject to regulations or exchange rules that often govern market making activities. Many institutional investors believe that due to the lack of regulation automated liquidity providers may simply withdraw from the market during crises, reducing liquidity at critical moments.

Finally, predatory trading refers to activities where a financial entity typically places thousands of simultaneous orders into a market while expecting to actually execute only a tiny fraction of the orders. Also known as market microstructure trading, this “place and cancel” process has two purposes. The first is an information gathering process. By observing which orders execute, the predatory trader expects to gain knowledge of the trading intentions of larger market participants such as institutional asset managers. Such asymmetric information can then be used advantageously in the placement of subsequent trades. A second and even more ambitious form of predatory trading is to place orders so as to artificially create abnormal trading volume or price trends in a particular security. The purpose is to intentionally mislead other traders and thereby gain advantage. Under the regulatory schemes of many countries there are general prohibitions against "market manipulation", but little if any action has been taken against predatory trading on this basis.

Therefore, algorithms implemented for high frequency trading have many uses and aim to profit from the markets using different approaches. Each of the trading types described above has a purpose and role to play in the financial markets, from market making, order routing to short term trend following and spotting arbitrage. Thus, although some strategies may seem unprofitable or place the trader in a disadvantageous position, there are motives and rewards behind such decisions.

Some practitioners believe automated trading puts the manual trading of retail investors, as well as institutional investors, in considerable disadvantage from a perspective of price discovery and liquidity (Arnuk and Saluzzi, 2008). A number of financial analytics/consulting companies including Quantitative Services Group LLC, Greenwich Associates and Themis Trading LLC have produced useful white papers on this topic.

2.2 Market Microstructure

Market microstructure, an expression coined by Garman (1976), is the study of how specific trading mechanisms affect the price formulation process. It covers the asset exchange process from placement and handling of orders to the translation into trades and transaction prices. Every market has procedures for matching buyers to sellers for trades to take place. In quote-driven markets dealers participate in every trade, whereas no intermediation is needed in order-driven markets where buyers and sellers directly trade with each other. One of the most critical questions in market microstructure concerns the process by which new information is assimilated and price formation takes place. Madhavan (2000) relates market microstructure to the study of how various frictions and departures from symmetric information affect the trading process. The following issues are studied in market microstructure theory:

- Market structure and design
- Price formation and price discovery
- Transaction and timing cost
- Information and disclosure
- Market marker and investor behaviour

The knowledge of market systems and structure is essential for a trader to decide in **which market** to trade and **when** to trade. Such knowledge also facilitates a trader in assessing the relative efficiency of the market and hence the arbitrage opportunities. In fact, the trading behaviour and trading costs are affected by market microstructure. This phenomena contests the theoretical predictions of classic arbitrage arguments that prices follow a random walk with drift (Samuelson, 1965), much like the beliefs of news analytics. In fact, these fields have parallel streams of thought where efficient markets and rational expectations are unproven in their theories. Our main focus is on "Information and disclosure" and "market maker and investor behaviour". More precisely, our interests lie in how markets react to information disclosed to traders, when and how this information is priced into assets, and how investors can benefit from these situations.

The models created to trade on market microstructure opportunities can be sorted into two types – inventory and information models. Similar to high frequency trading, transaction costs are a vital factor in profiting from market microstructure trading, which is why this is a trading strategy commonly adopted by HFTs. Profits earned are marginal with gross average gain for a position held for only a few seconds being a few basis points (1bp=0.01%). However, this varies according to different types of traders for example, traders on a proprietary trading desk face 1bp or less of transaction costs whereas other institutional players such as hedge funds can expect to be charged from 3bps to 30bps.

Inventory models are in fact the liquidity provision models previously mentioned. By satisfying dealer book imbalances price changes are induced through order flow, giving this type of trading another name known as market making. Liquidity traders craft order submission strategies by following short-term price momentum and aim to profit from providing liquidity. Although they have little proprietary information about the true value of the security, the shape of the order book can show impending changes in market price. Hence with such predictive indications market making traders actively exploit this.

Information models address the intent and future actions of various market participants by trading on information flow and possible informational asymmetries arising during the dissemination of said information. Different types of market participants receive news through different resources, for example market makers obtain their information from what is conveyed through order flow and changes in the bid-ask spread is the supply for other market participants. Asymmetric information flow does exist with institutional traders being positioned disadvantageously. Their lack of information regarding time-varying market liquidity erodes their speculative profits (Hong and Rady, 2002). The main outcome of such models is that bid-ask spreads still persist even when the market maker has unlimited inventory and any trading request is instantaneously absorbed. Brennan and Subrahmanyam (1996) specify a vector autoregressive (VAR) model for estimation of an information-based impact measure.

Harris (1998) identifies three types of traders:

(i) Informed investors
(ii) Uninformed investors (Liquidity traders)
(iii) Value-motivated traders

Informed investors are those who possess material about an impending market move and consequently can influence the order flow of stocks in a biased manner. High frequency money managers often fill this role and tend to execute their orders close to or at market prices, utilising limit orders more than liquidity traders (Bloomfield, O’Hara and Saar, 2005). They are able to assimilate all available information about a given stock and thereby reach some certainty about the market price of the stock. Such information may be acquired by subscription to (or purchased from) news sources; typically FT, Bloomberg, Dow Jones, or Thomson Reuters. They might have access to superior predictive analysis which enhances their information base. Inversely, uninformed investors are those who are not aware of any information affecting prices besides order volume. Therefore, in order to earn excess returns, uninformed traders follow a strategy of providing liquidity and tracking short-term price momentums thus acquiring the alternative names of liquidity traders or inventory traders (O’Hara, 1995). In fact, this form of liquidity provision is instrumental to the markets. By simply keeping an inventory of stocks, liquidity traders are market making and causing less friction in the trading of stocks whilst realising marginal gains through the use of limit orders, often many times intraday. Individual traders and retail traders are uninformed investors, also known as noise traders. The third form of traders are those who wait for security prices to become cheap relative to their valuations based on fundamental indicators, known as value-motivated traders. They apply predictive analytic models and use information to identify trading opportunities by spotting inefficiencies and mispricing of stocks. Although uninformed investors and value-motivated traders do not have access to exogenous information, what they are able to exploit is the value of such information, often extracted from anticipated announcements and then applied in predictive pricing models. Bloomfield, O’Hara and Saar (2005) investigate the evolution of liquidity in an electronic limit order market where informed traders, liquidity traders and salient features of electronic limit order markets exist. Their main result concludes that liquidity provision shifts as trading progresses, with informed traders increasingly providing liquidity in markets. In quote-driven markets, limit orders are the major provider in liquidity.

An additional role that especially needs to be mentioned is that of the market maker. Their responsibility is to set the price for the bid quote and ask quote and stand willing to buy or sell securities on demand. Thus these market participants control the liquidity of securities, particularly in the case of order-driven markets. In order to meet the supply and demand of traders, market makers need information with regards to market conditions to offset any suboptimal positions and information asymmetry. The objective is to achieve a rapid inventory turnover and not accumulate significant positions on one side of the market. The bid-ask spread is the profit that is gained. According to Bagehot (1971), market makers compensate themselves for bad trades due to the adverse selection of insiders by making the market less liquid.

2.3 Liquidity: Measures and Implications

Traditionally literature has focused on the estimation of stock price return and volatility in a trading context and to a certain extent fund management. In the field of news sentiment, this is also the case with the majority of studies researching the relationship between news and return or volatility. However, after several incidents on the stock market in recent years (e.g. Flash crash of 6 May 2010), the importance of liquidity and the risks entailed to trading has come to light. The availability of liquidity to traders and brokers is now so crucial that regulators have sanctioned new rules forcing stricter requirements on institutions. Henceforth, research on news sentiment and its implications on liquidity have been explored (Gross-Klaussman and Hautsch, 2011; Smales, 2013; Riordan et al., 2013).

Liquidity has many definitions depending on which perspective it is being observed and measured. Generally, liquidity refers to the readiness by which an asset can be converted into cash (or vice versa) at the lowest possible transaction cost. Transaction costs include both explicit costs, e.g. brokerage, taxes, and implicit costs, e.g. bid-ask spreads or market impact costs. A market is termed liquid when traders can trade without significant adverse effect on price (Harris, 2002). Measures of liquidity can initially be separated into two broad categories: trade-based measures and order-based measures (Aitken and Comerton-Forde, 2003). Trade-based measures include trading value, trading volume, trading frequency, and the turnover ratio. These measures are mostly ex-post measures. DeLong et al. (1990) predict an increase in trading volume when absolute values of sentiment are high due to the behaviour of liquidity traders. High absolute values of sentiment induce high trading volume due to increased trading activity from liquidity traders and market makers trying to maintain equilibrium in the markets. Theory suggests that informed trading after information events should increase (Kim and Verrecchia, 1994).

Order-based measures have many forms and are more commonly used due to their more informative nature. They are classified into the following three groups:

1. Tightness measures: the ability to buy and sell a stock immediately.
2. Depth measures: the ability of the market to process large volumes of trade without affecting the current market price.
3. Resiliency: the ability of the market to return to its “normal” level after absorbing a large order.

Tightness, also known as spread, measures provide a clear indication of the costs associated with transacting and are represented by the bid-ask spread and its many variations. A decrease in spread measures indicates an increase in liquidity. The bid-ask spread is calculated as:

$$Spread_t = \frac{p_t^A - p_t^B}{p_t^M} \qquad [2.1]$$

where p_t^A is the ask price at time t, p_t^B is the bid price at time t, and p_t^M is the mid-price between the bid and ask price at time *t.* In fact, this measure is more accurately known as the relative spread because it allows direct comparison of spreads for different assets. Variations of this include using log prices or the last paid price of the most recent trade.

A signal of incoming news is the widening of bid-ask spreads. Liquidity suppliers react to news by reducing order aggressiveness in revising quotes to avoid the costs incurred by trading with informed traders. Similar to stock price returns, news sentiment has been found to have an asymmetric impact on spread measures but a consistent increase around news arrival for market depth (Riordan et al., 2013). In the presence of (extreme) sentiment, liquidity should also increase for highly traded stocks i.e. spreads decrease, as theory suggests that sentiment leads to increased trading activity. The adverse selection component of the bid-ask spread can be interpreted as private information that is impounded into prices through trading.

Effective spread is a trade process based measure of liquidity that is calculated as the spread paid when an incoming market order trades against a limit order. The equation for it is:

$$EffSp_t = |p_t - p_t^M| \qquad [2.2]$$

where p_t denotes the last traded price before time t and p_t^M is the mid-price calculated as above. A general rule of thumb is taken to be: if the effective spread is smaller than half the absolute spread, then this reflects trading within the quotes (see Chordia, Roll and Subrahmanyam, 2000 or Hasbrouck and Seppi, 2001). Several manipulations are added to this measure to enhance its comparability with other spread measures, i.e. multiplying by two or weighting with the number of trades.

Depth measures are a sign of illiquidity indicating an adverse market impact for investors. Moreover, these measures regard the volume at the best bid and ask prices. A commonly used measure of market depth is called Kyle's Lambda (Kyle, 1985):

$$\lambda = \frac{r_t}{NOF_t} \qquad [2.3]$$

where r_t is the asset return and NOF_t is the net order flow over time. The parameter λ can be obtained by regressing asset return on net order flow over some time window. When time *t* is chosen to be very short, such as tick trades, λ reduces to the simple equation above. A highly liquid stock would incur small changes in price i.e. little return, for a given level of trading volume. This measure can also be used to calculate market impact. Kyle (1985) states that the prices determined by market makers are assumed to equal the expectation of the liquidation values of the commodity, conditional on the market makers' information sets at the dates the prices are determined.

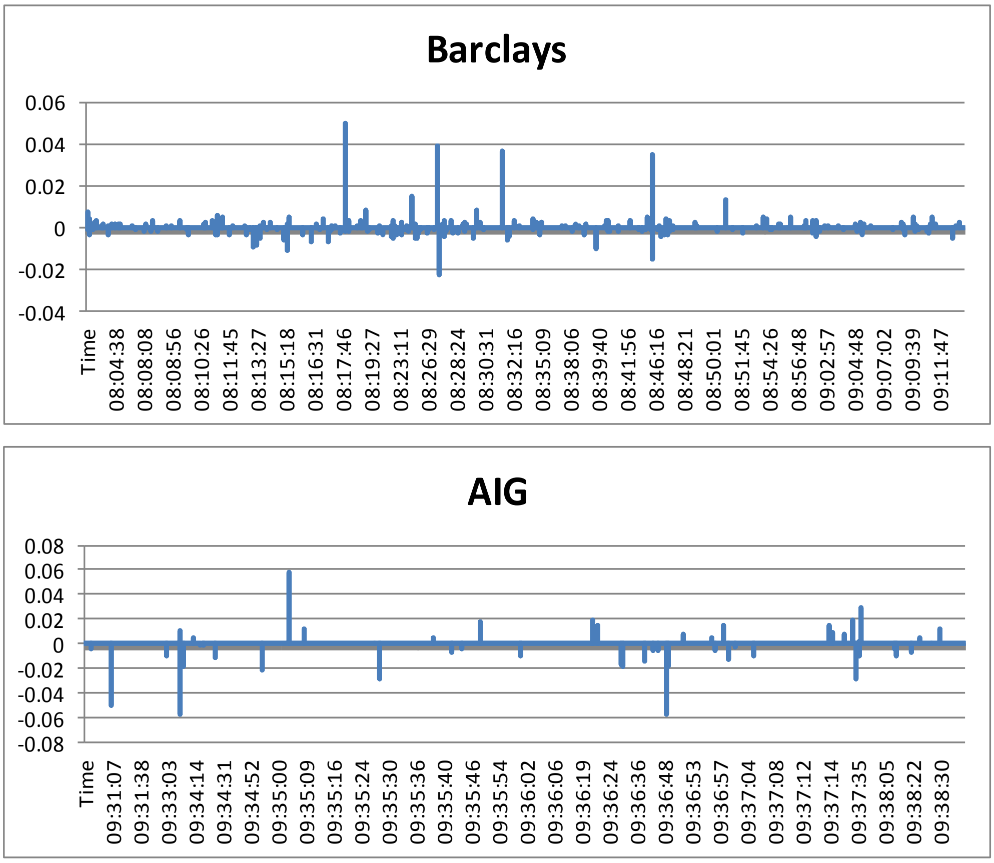


Figure 2.3: Kyle's λ values calculated for all trades on Barclays and AIG in the day 1 June 2010.

Another measure of market depth is Hui-Heubel (HH) liquidity ratio (Hui and Heubel, 1984). This model was used to study asset liquidity on several major U.S equity markets, and relates trading volume to the change of asset price. Given the market activities observed over N unit time windows, the maximum price P_{max}, minimum price P_{min}, average unit closing price $\bar{P}$, total dollar trading volume V, and total number of outstanding quotes Q, the Hui-Heubel L_{HH} liquidity ratio is given as follows:

$$L_{HH} = \frac{(P_{max} - P_{min})}{\frac{V}{Q} * \bar{P}} \qquad [2.4]$$

A higher HH ratio indicates higher price to volume sensitivity. Additional measures include market depth, trading volume and more.

The resilience dimension of liquidity refers to the speed at which the price fluctuations resulting from trades are dissipated. Market-efficient coefficient (MEC) (Hasbrouck and Schwartz, 1988) uses the second moment of price movement to explain the effect of information impact on the market. If an asset is resilient, the asset price should have a more continuous movement and thus low volatility caused by trading. Market-efficient coefficient compares the short term volatility with its long term counterpart. Formally:

$$MEC = \frac{Var(R_{long})}{T * Var(R_{short})} \qquad [2.5]$$

where T is the number of short periods in each long period. A resilient asset should have a MEC ratio close to 1. Calculating the MEC for a handful of assets from the US and UK stock markets it can be shown that US assets are more resilient than UK assets (Figure 2.4).

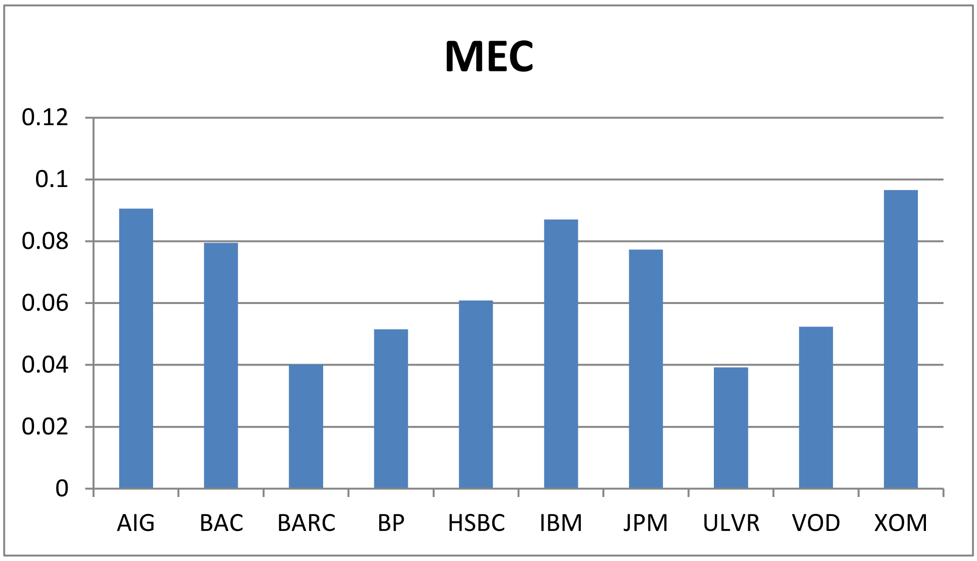


Figure 2.4: The market-efficient coefficient (MEC) calculated for a handful of US and UK stocks.

Alternative measures for resilience include intraday returns, the variance ratio and the liquidity ratio.

Literature also has precedence for another aspect of liquidity: *immediacy* - the speed at which trades can be arranged at a given cost. Illiquidity can be measured by the cost of immediate execution (Amihud and Mendelson, 1986). Thus, a natural measure of illiquidity is the spread between the bid and the ask prices. Later, Amihud (2002) modified the definition of illiquidity. The now-famous illiquidity measure is the daily ratio of absolute stock return to its dollar volume averaged over some period:

$$ILLIQ_{iy} = \frac{1}{D_{iy}} \sum_{d=1}^{D_{iy}} \frac{|R_{iyd}|}{VOLD_{iyd}} \qquad [2.6]$$

where R_{iyd} is the return on stock *i* on day *d* of year *y* and $VOLD_{iyd}$ is the respective daily volume in dollars. D_{iy} is the number of days for which data are available for stock *i* in year *y*.

The vast literature on liquidity studies the relationships of liquidity and the cost of liquidity with various stock performance measures, trading mechanisms, order-trader types and asset pricing. Acharya and Pederson (2005) present a simple theoretical model (liquidity-adjusted capital asset pricing model- LCAPM) that helps to explain how liquidity risk and commonality in liquidity affect asset prices. The concept of commonality of liquidity was highlighted by Chordia et al. (2000) when the authors stated that liquidity is not just a stock-specific attribute given the evidence that the individual liquidity measures (quoted spreads, quoted depth and effective spreads) co-move with each other. Later Hasbrouck and Seppi (2001) examine the extent and role of cross-firm common factors in returns, order flows, and market liquidity, using the analysis for 30 Dow Jones stocks. According to Hasbrouck (2006), the Amihud illiquidity measure is most strongly correlated with the TAQ-based price impact coefficient among the daily proxies.

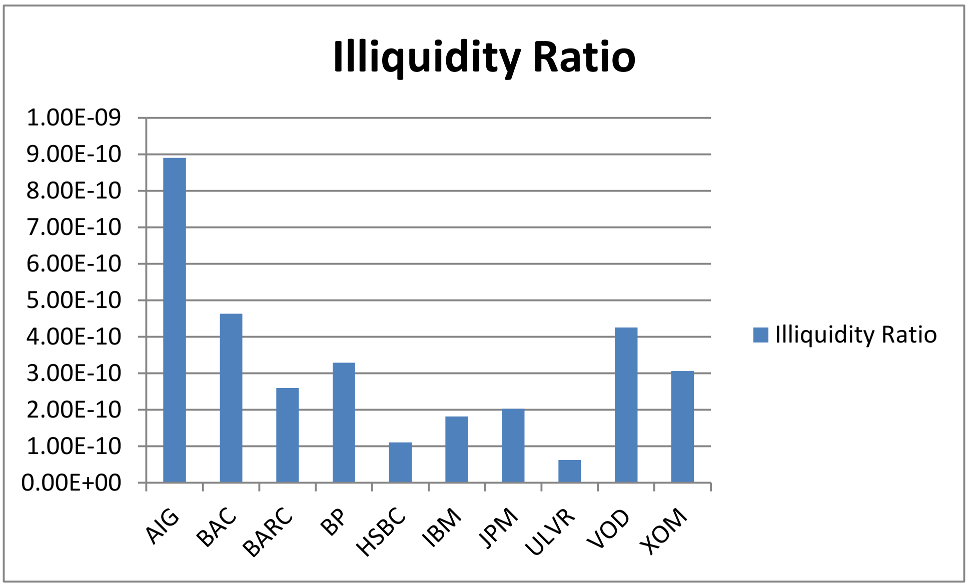


Figure 2.5: Illiquidity ratios for a handful of US and UK stocks.

A recent study of news sentiment on intraday price discovery, liquidity and trading intensity by Riordan et al. (2013) confirm the increase of adverse selection costs, liquidity and trading intensity around news. Furthermore, an asymmetric effect was found between negative news and positive news (neutral news caused similar reactions as positive news), with the former being associated more with higher adverse selection costs and lower liquidity. This is further amplified in periods before release of negative news.

Many studies have found significant relationships between news events and liquidity, with Krinsky and Lee (1996) leading the field by analysing the change in spread around scheduled earnings announcements. Their results show an increase in the adverse selection component of the spread during announcements, which they attribute to information advantages of informed traders and faster news processing capabilities of public information processors. Ranaldo (2008) examines firm specific unstructured news at the Paris Bourse and finds an increase in liquidity as well as higher adverse selection costs around news release. Furthermore, Chordia, Roll and Subrahmanyam (2001) find that daily changes in market averages of liquidity are highly volatile and there are strong days-of-the-week effects. Also, in down markets, liquidity tends to plummet. Baker and Stein (2004) even suggest liquidity as a type of sentiment index, in the form of share turnover, and argue that signs of high liquidity are a symptom of overvaluation.

To the extent that a broad market or a particular security becomes more volatile, it can be expected that liquidity providers will demand greater compensation for risk by widening bid-ask spreads. This is confirmed in recent research reported by Gross-Klussmann and Hautsch (2011) who conclude that by capturing dynamics and cross-dependencies in the vector autoregressive modelling framework they find the strongest effect in volatility and cumulative trading volumes.

Chapter 3

News Sentiment and Its Market Impact

3.1 News Metadata

News in some sense is not an exact term; thus news can be associated with many different types of information. We provide below Leinweber's (2009) broad classifications to distinguish the different forms.

1. News This refers to mainstream media and comprises the news stories produced by reputable sources. These are broadcast via newspapers, radio and television. They are also delivered to traders' desks on newswire services. Online versions of newspapers are also growing in volume and number. News may be separated into two categories: scheduled and unscheduled. Macroeconomic announcements are an example of scheduled news where the time factor is known yet the content is unknown. On the other hand, unscheduled news possesses undetermined factors for both time and content.

2. Pre-news This refers to the data source that reporters research before they write news articles. It comes from primary information sources such as Securities and Exchange Commission reports and filings, court documents and government agencies. It also includes scheduled announcements such as macroeconomic news, industry statistics, company earnings reports and other corporate news.

3. Web 2.0 and social media These are blogs and websites that broadcast ''news'' and are less reputable than news and pre-news sources. The quality of these varies significantly. Some may be blogs associated with highly reputable news providers and reporters (for example, the blog of BBC's Robert Peston). On the other hand, some blogs may lack substance and be fuelled entirely by rumours. Social media websites fall at the lowest end of the reputation scale. Barriers to entry are extremely low and the ability to publish ''information'' is easy. These can be dangerously inaccurate sources of information. At a minimum they may help us to identify future volatility. Individual investors pay relatively more attention to the previous two sources of news than institutional investors. Information from the web may be less reliable than mainstream news. However, there may be ‘‘collective intelligence’’ information to be gleaned. That is, if a large group of people have no ulterior motives, then their collective opinion may be useful (Leinweber, 2009, Ch. 10, Preis, Moat and Stanley, 2013).

The news metadata that we study is compiled from the first category of news. News content vendors collect news from a range of sources worldwide such as electronic newswires, newspapers and magazines. Real time newsfeeds are predominantly supplied by newswires. They employ journalists from all over the world whose reports are then distributed to other organizations in the industry. In order to expand their coverage on a global scale, translation of headlines and text is obligatory. Several complications may occur during this process that will deter the accuracy of sentiment scoring later on, which is why the current coverage of news in data does not span outside the English language. Progress however is being made in this area to guarantee the availability of first-hand information.

News analytics data is presented in a metadata format, which is a term promoted by the increasing popular field of Big Data. Metadata refers to information that describes a set of data and is highly applicable in understanding the material stored in data warehouses. Essentially it is “data about data”, incorporating details of how, when and by whom the data was collected and how the data is formatted. Furthermore, its importance in XML-based web applications cannot be overlooked as webpage content is commonly described using meta tags, which consequently is how search engines determine their search index. Applying a similar technique, data vendors have also constructed machine learning algorithms to identify relevant tags for a news story. These tags turn the unstructured stories into a basic machine readable form and are often stored in XML format. They reveal the story’s topic areas and other important metadata, such as a list of companies the story includes in its writing. With the majority of news available electronically and online, it is appropriate that the processing of news is in metadata format. The characteristics given as data fields to news metadata include (i) relevance, (ii) novelty and (iii) sentiment scores based on an individual asset (but this is not an exhaustive list). Thus the analytical process of producing such scores is fully automated from collection, extraction, aggregation, to categorisation and scoring. The result is an individual score assigned to each news article for each characteristic using scales from 0-100 or probabilities.

For our research investigations, news metadata from both RavenPack and Thomson Reuters were available at our disposal. Experiments were carried out on both sets of data but we do not set out to make any comparisons. Some of the descriptions and knowledge of data fields explained below were taken from their respective manuals for news analytics data (RavenPack, 2010 and Thomson Reuters News Analytics, 2010). Both sets of data are similar in structure (see Appendix A and B). In our study, we have used only Thomson Reuters news metadata.

A news metadata record may relate to one or more companies (assets). Thus one news article may produce a set of attributes (scores) for more than one company. Similarly, an individual asset may be linked to a repeated series of news (events) on a given theme. This leads to the concept of novelty of "news items" in the series. Therefore, metrics that can distinguish between this information is required, namely, the data fields known as relevance, novelty, volume and headline classification to name a few.

The primary metadata fields, as with all data, are **date and time** (including time zone). The general format used is "DD MM YYY hh:mm:ss:sss" where time-stamps are accurate to the nearest millisecond. These fields are necessary for archive purposes. Next on the list is **company ID**, where it may be presented as the universal ticker symbol or the data vendor's unique coding.

Relevance is defined as the significance of one news article on a particular asset. It is normally measured as a score between 0 and 100. This is an important factor in the filtering of news metadata as the difference between a news event's key role and the name of a passing by competitor is rated by relevance. Setting a high benchmark for relevance score guarantees a significant linkage between the news item and the selected company.

Novelty measures the originality of a news article. Also scored between the range 0-100, it is calculated through comparison with previous news items containing the same asset name. Linguistic comparison programs are employed to identify similarity counts between articles and complete novelty is reached once a particular threshold is surpassed. Novelty scores are provided for a number of time periods depending on the frequency that is needed, i.e. 12 hours, 24 hours, 3 days etc. Moreover, a linked count is also provided, relating that particular news item to the original article. Therefore, if the item scores lower than 100 in the novelty field then it will have a linked count greater than zero, with larger numbers representing slower updates.

Headline category is an exhaustive list of possible topics that the news item could be regarding. This data field is useful for segregation of news types, e.g. scheduled news, macroeconomic announcements.

The news metadata therefore can be effectively filtered (extracted) by specifying limits on many of these attributes. The focus of our work is concentrated on non-scheduled news and the surprise effect on markets. Hence, any social media and data from the web are not considered. The attributes of news stories used in our study are Relevance and Sentiment, where a benchmark score of 70 was chosen for relevance. The scores used to denote sentiment are described below in section 3.2. Although we use intraday news metadata, for a given asset the number of news stories, hence data points are variable and do not match with the time frequencies of market data.

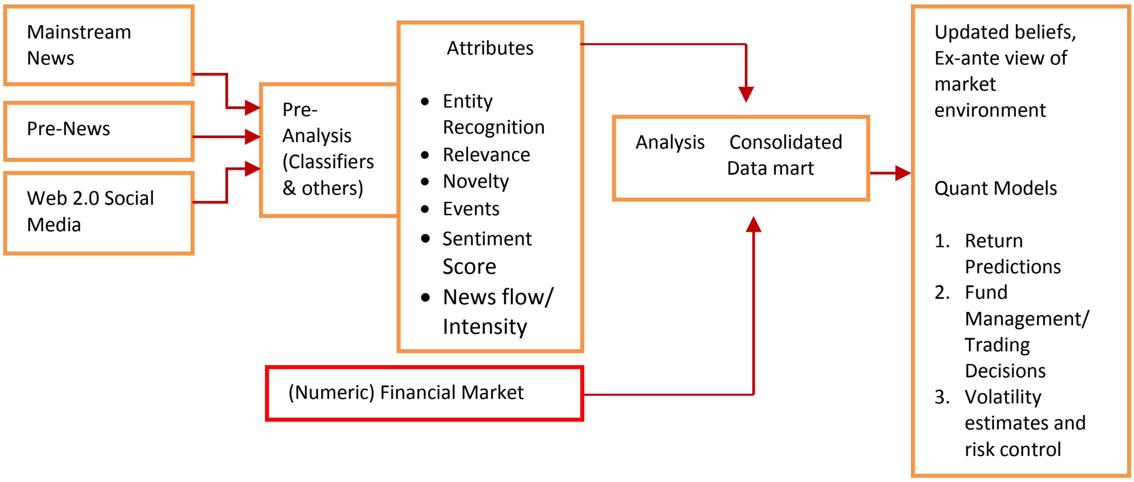


Figure 3.1: An outline of information flow and modelling architecture of news metadata.

3.2 Sentiment score

A news sentiment score measures the emotional tone within a news item and varies between positive and negative. A sentiment score can be defined as follows: a value falling within a range consisting of a minimum and maximum depicting the overall tone of a news article. Depending on the measurement of scale, the exact polarity of sentiment in the news can be deduced, i.e. Thomson Reuters assign probabilities to the moods “Positive”, “Neutral” and “Negative” to infer an overall sentiment that is the average of all 3 scores, whereas RavenPack directly produce a sentiment score belonging to the range 0-100 that then allows a conclusion of positivity or negativity. How these scores are derived is a process known as sentiment classification and is described fully in section 1.4.

By automating the judgement process, the human decision maker can act on a larger, hence more diversified, collection of assets. These decisions are also taken more promptly thus reducing the latency of trades. The intuition is that somewhere within these series of news sentiment metadata lies indicative signals waiting to be discovered and revealed. Through experimentation with Reuters’ sentiment data, Uhl (2011) concludes that there is a clear advantage in disregarding neutral sentiment as such ambiguous text blurs the overall mood. This conclusion was based on a better performance by trading strategies that only considered positive or negative scores and not the whole population. Similarly with the same set of data, Sinha (2010) tests the predictability of asset returns using a qualitative information measure that is based on news sentiment scores. The measure is able to predict weekly returns and mitigates short-term reversal in the weekly momentum strategy. Another important result concluded by researchers on the sentiment of news events is that more emotive news (highly positive and highly negative) can better predict volume and volatility increases, more so by the negative extremes. Thus, the points raised above demonstrate an encouraging effect to be explored between news sentiment and asset behavioural characteristics, and also provides motivation for our study into the connection of news sentiment and return, volatility and liquidity.

Thomson Reuters’ news sentiment engine analyses and processes each news story that arrives as a machine readable text. Through text analysis and other classification schemes the engine then computes values for the attributes described in section 3.1.

As already explained, a news event sentiment can be positive, neutral and negative and the classifier assigns probabilities such that

$$Prob(positive) + Prob(neutral) + Prob(negative) = 1.0 \qquad [3.1]$$

We turn these three probabilities into a single sentiment score in the range 0-100 using the following equation:

$$\hat{S}Sent = 100 * (Prob(positive) + \frac{1}{2}Prob(neutral)) \qquad [3.2]$$

where $\hat{S}Sent$ denotes a single transformed sentiment score. Although the probability of negative sentiment cannot be seen in equation 3.2, it is however implied from equation 3.1 and thus is considered in the calculation of $\hat{S}Sent$. Some examples of this numerical conversion are provided below.

POSITIVE	NEUTRAL	NEGATIVE	$\boldsymbol{\hat{S}Sent}$
1	0	0	**100**
0	0	1	**0**
0.1	0.05	0.85	**12.5**
0.9	0	0.1	**90**

Table 3.1: Examples of how the sentiment score $\hat{S}Sent$ is constructed.

We find that such a derived single score provides a relatively better interpretation of the mood of the news item. Thus the news sentiment score is a relative number which describes the degree of positivity and negativity in a piece of news. We further shift the sentiment score by subtracting 50 (the value corresponding to the neutral sentiment score = 0) to compute

$$SENT = \hat{S}Sent - 50 \qquad [3.3]$$

Thus the score *SENT* lies between -50 and +50. During the trading day, as news arrives it is given a sentiment value.

To test for the existence of a relationship between news sentiment scores and respective asset prices, a χ^2 test is carried out. Due to the arrival of news at different times asynchronously, it is not possible to construct a corresponding time series of asset price data. Therefore, a method is taken to test for independence between sentiment scores and the change in prices, specifically the direction of change in prices. Our interest lies in whether negative (positive) news sentiment correspond to price movement downwards (upwards) in the next time period, thereby focusing on the reaction of asset prices to news sentiment. Preliminary observations can be made from the contingency table where both variables are sorted into quintiles (see Table 3.1). A clear observation is that the majority of data points fall within the third quintile (i.e. median) of price change and the upper and lower quintiles of sentiment scores. In other words, both highly positive and highly negative news sentiment frequently correspond to little or no movement in stock prices, as the third quintile bounds zero. Taking the null hypothesis to be that the two variables are independent, we calculate a test statistic of 40.77. Testing at the 1% level we reject the null hypothesis and conclude that causality between these two variables could exist. However, it should be noted that this test statistic is only slightly larger than the critical value of 39.25 and produces a p-value of $5.98e^{-4}$. As a consequence, it has been established that these two variables are not independent whilst at the same time, causality cannot be confirmed. It is well known to practitioners that news flow affects prices, we have therefore derived a new measure, the impact score, which takes into account news volume (see section 3.4).

	I1	I2	I3	I4	I5
C1	37	16	8	59	16
C2	54	23	10	41	17
C3	1031	586	184	906	577
C4	64	18	10	44	18
C5	33	21	7	53	25

Table 3.2: Contingency table for price change of AIG in 2008 and its respective sentiment score, split into quintiles. I1-I5 represent the sentiment score quintiles and C1-C5 represent the quintiles of price change.

3.3 News flow

A characteristic of news events not included in the metadata is the volume of news for a particular asset. In literature this is commonly referred to as news flow, embodying the rise and fall of news counts over a specific period. As a consequence, initial start dates must be stated in order to consider news flow in a comparable manner. Intuitively such a factor would influence the degree of impact of news sentiment on stock markets. Consider the case of a single news item that reports of faulty products being distributed by a particular company, and compare it to an instance where all major newswires are reporting delays in meeting customer demand by another company. For both cases, the news is bad with possible damaging implications on the companies, and so the sentiment can easily be deduced to be negative. However, an evident difference exists, that is, the magnitude of the negative sentiment, which consequently will affect the subsequent reaction by the markets. The single news report will portray negativity as an overall sentiment; however the impact will be small in comparison to bad information that is broadcasted by many news sources.

Stock prices can be impacted purely by the frequency of news events (news flow) as research by Tetlock (2011) proves. He discovers a reaction by investors to news that is already considered stale, causing temporary movements in stock prices. Thus, information which is no longer novel and has been reported several times can still affect investors' trading decisions, implying that volume of news can enhance the longevity of information and emphasize sentiment polarity. Similarly, Dzielinski, Rieger and Talsepp (2011) relate a high concentration of news to volatility asymmetry. The reasoning is as follows: bad news is typically prevalent in times of high news concentration, and private investors are known to react nervously to bad news (Talsepp and Rieger, 2009), hence the overreaction of private investors to bad news will likely lead to the observed asymmetry in volatility.

Other features found to be linked to news flow include size of companies. Larger companies (with more liquid stock) tend to have higher news coverage/news flow. Moniz, Brar, and Davis (2009) observe that the top quintile of company sizes accounts for 40% of all news articles and the bottom quintile for only 5%. Cahan, Jussa, and Luo (2009) also find news coverage is higher for larger market capitalisation companies as well as determining a correlation around 55% between news flow and market capitalisation.

The justification for aggregation of news sentiment is related to news flow. The influential power of many articles on one news story exceeds that of a single article. Hence the combined sentiment will indicate to investors that this is more trustworthy information which could impact the markets due to a stronger score. The differing number of news covering a story is known as news intensity.

In consideration of the arguments provided above and through results produced in the existing literature, we believe that it is necessary to incorporate the news flow factor into the study of news sentiment effects on asset behaviour. Thus, we extract the necessary metadata and calculate the news flow. Therefore, using this as a motivation we construct a novel measure called the Impact score. As we will show in the next section, this additional feature of the sentiment score will enhance the relationship between news sentiment and asset prices as well as offering a more realistic interpretation of news events.

3.4 Impact score

In this thesis we have considered two measures of news sentiment. The first is sentiment score, which is explained in section 3.2. To date all the studies that report on news and asset behaviour only consider sentiment (see section 3.2). We construct a new measure called impact score which we find more relevant and useful in predicting such asset characteristics. The sentiment score only quantifies the mood of a typical investor in respect of a news event. The impact score, on the other hand, takes into consideration the decay of the sentiment of one or more news events and how after aggregation these impact the asset price behaviour.

It is well known from research studies that news flow affects asset behaviour (Patton and Verardo, 2012, Mitra, Mitra and diBartolomeo, 2009). Therefore, accumulation of news items as they arrive is important. Patton and Verardo (2012) noticed decay in the impact of news on asset prices and their betas on a daily timescale and further determine the complete disappearance of news effects within 2-5 days. A similar effect was also observed in Arbex-Valle et al. (2013). Mitra, Mitra and diBartolomeo (2009) created a composite sentiment score in their volatility models and after initial experiments saw no effects to volatility predictions with sentiment alone. Their decay period was over 7 days.

So as recent literature suggests, not only does it produce better results but it is also more realistic to extend the influencing duration of news on assets, lasting longer than the brief period of news release. Bearing this in mind, we form a new measure of news sentiment which better incorporates the impact of news events by amalgamating the attenuation and decay of sentiment scores for a particular asset. This is named the Impact score.

A component within the news sentiment metadata provided by RavenPack is called News Impact Projection (NIP). It takes into account the price impact of stocks mentioned in a news story headline. The formation of this impact score utilises several advanced machine learning techniques with the common objective of identifying the probability of a particular stock’s volatility to be higher or lower than the volatility of the market.

In order to compute the impact of news events over time we first find an expression that describes the attenuation of the news sentiment score. The impact of a news item does not solely have an effect on the markets at the time of release but also over finite periods of time that follow. To account for this prolonged impact, we have applied an attenuation technique to reflect the instantaneous impact of news releases and the decay of impact over a subsequent period of time. The technique combines exponential decay and accumulation of the sentiment score over the time buckets under observation. We take into consideration the attenuation of positive sentiment to the neutral value and the rise of negative sentiment also to the neutral value and accumulate (sum) these sentiment scores separately (see equation 3.4 and 3.5). By separating the decay process between positive and negative sentiment scores we avoid an exact cancellation to zero, which may be interpreted as no news.

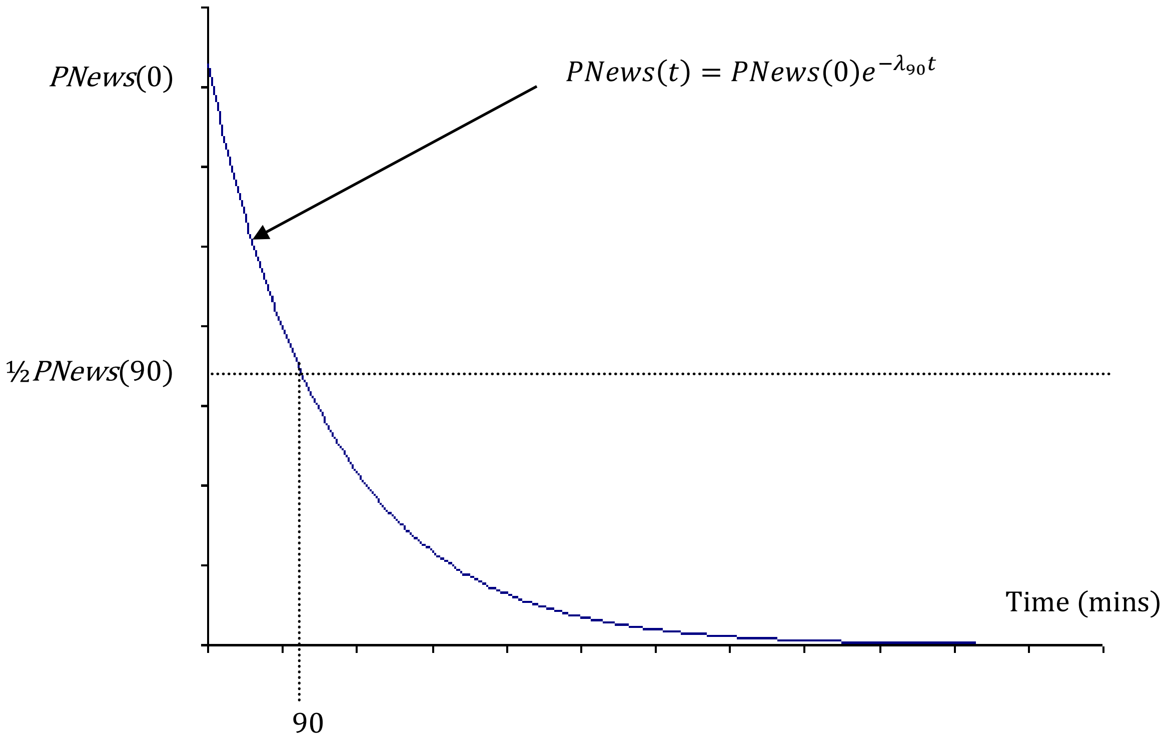


Figure 3.2: *A representation of news sentiment decay.*

Figure 3.2 shows the decay rate of news sentiment. The parameter λ_{90} specifies that the news sentiment decays exponentially to half of its initial value over a 90 minute period and is the chosen value for our experiments. For the rest of this thesis, λ_{90} will be expressed as λ.

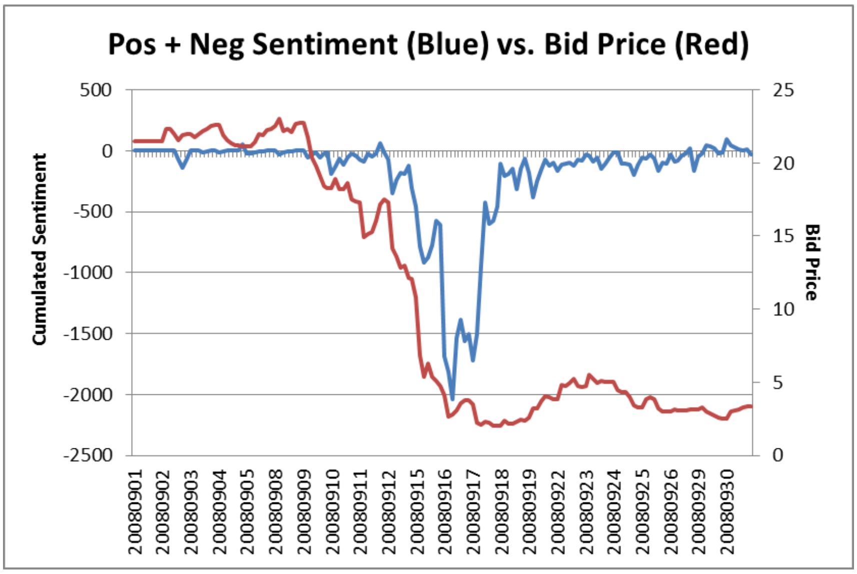


Figure 3.3: The bid price and cumulated (positive + negative) sentiment scores for AIG for August 2008.

The sum of the sentiments for different news arrivals in a given time bucket of one minute duration (that is, 630 buckets during the trading day) gives the overall sentiment in that time bucket. Therefore, if no news arrives within a certain time bucket then no sentiment is added on, leaving the further decayed value from the previous time bucket as the sentiment score. In Figure 3.3 we plot the overall sentiment for the asset AIG over one month of data in which the vertical axis on the left is the sentiment score, the vertical axis on the right is the bid price and the horizontal axis is 630 x 21 = 13230 minute-bar time buckets.

The work of Leinweber and Sisk (2011) and also earlier work by Mitra, Mitra and diBartolomeo (2009) as well as the present study indicate that accumulation of sentiment leads to a fairly good fit with asset price in general and with asset price volatility in particular.

Data Granularity:

Recall from section 2.1 that our interests lie in studying the effects of news sentiment intraday. Uses for such work include establishing trading strategies or quantifying risk. Therefore, our data granularity is set to be minute bar.

The **trading day** starts at 08:00 hours and ends at 18:30 hours thus in a trading day the total number of buckets is 630.

Any news arriving overnight or during the weekend is bucketed into the next morning or the day's first minute, where the size of a bucket is 1 minute. Hence, the assumption is taken that the impact of such overnight and weekend news is incorporated into prices the following day.

Although news arrives asynchronously we work out the aggregated impact of all news in the following way.

Let

POS denote the set of news with positive sentiment value $SENT$>0;
NEG denote the set of news with negative sentiment value $SENT$<0;

and

$PNews(k, t_k)$ denotes the sentiment value of the k^{th} positive news arriving at time bucket t_k, 1≤ t_k ≤630 and $k \in POS$; $PNews(k, t_k) > 0$.
$NNews(k, t_k)$ denotes the sentiment value of the k^{th} negative news arriving at time bucket t_k, 1≤ t_k ≤630 and $k \in NEG$; $NNews(k, t_k) < 0$.

Let λ denote the exponent which determines the decay rate. We have chosen λ such that the sentiment value decays to half the initial value in a 90 minute time span. The justification for these values is as follows. Firstly, it should be clarified that there are two variables to be determined – the rate of decay (speed) and the decay duration (how long the decay for one piece of news lasts). Next, each variable was determined through several variations. To establish the appropriate value for the decay duration, a default value for the decay rate was initially taken to be a half. In consideration of the fact that the market data is of minute-bar frequency, an adequate timeframe of 15 minutes was chosen as the starting point for testing. Continuing in this manner, durations were extended to test at 30 minutes, 60 minutes and 90 minutes (see Figure 3.4).

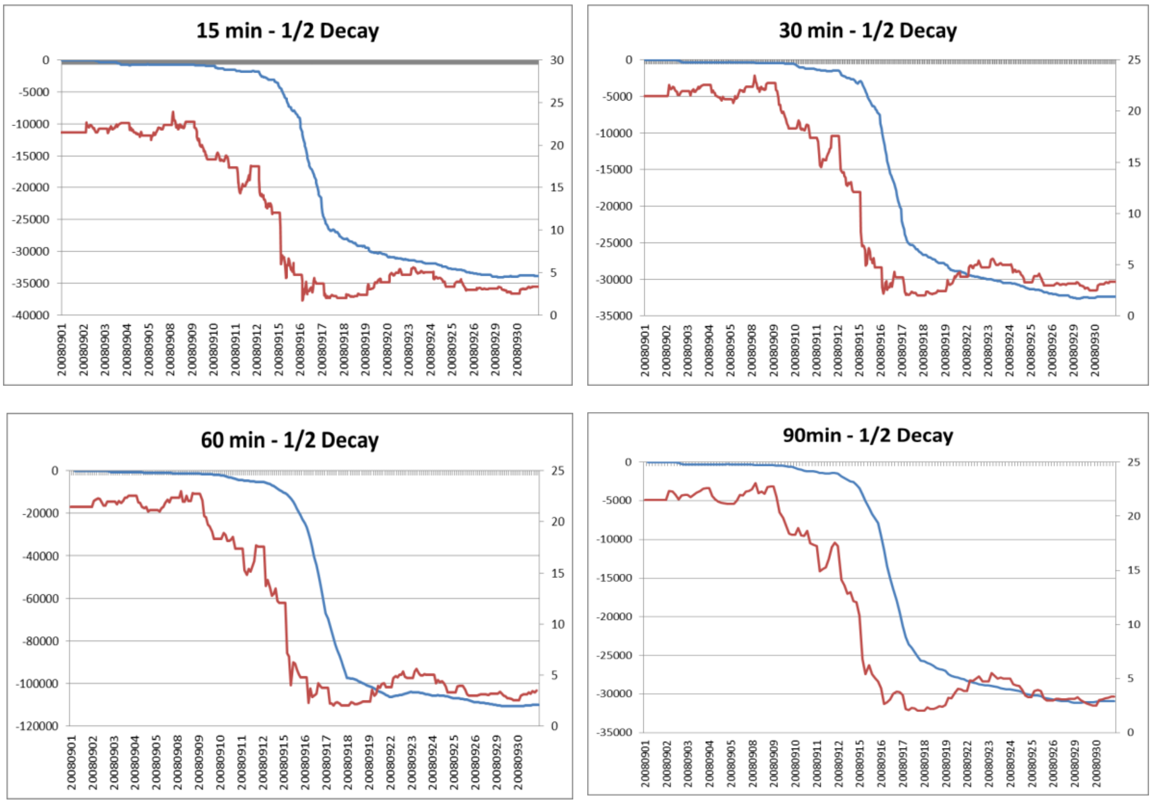


Figure 3.4: The half-life decay of sentiment scores for AIG over the month of September 2008 combined with different decay durations of 15 minutes, 30 minutes and 60 minutes respectively. The red line depicts the bid price.

With regards to the decay rate, half-life decay was taken as the starting point to construct the impact score. As a robustness check, other decay rates were also considered such as two-thirds and three-quarters, however, results concluded that the decay rate of half is most suitable. The justification of such conclusions can be seen from Figure 3.5, which plots the different combination of decay rates with the bid price (decay duration is held constant at 90 minutes). It can be observed that the movement of cumulated sentiment scores is most synchronised with the bid price movement for the decay rate of one half. This is most prominent in the drop of sentiment and prices during the days 10-16 August 2008, where the trough of the cumulated sentiment score corresponds to the lowest price. For the other larger decay rates, there is a mismatch in this movement, although it is small for the case of decay rate three-quarters.

Thus, the cumulated positive and negative sentiment scores for one day are given by equations 3.4 and 3.5.

$$PImpact(t) = \sum_{\substack{k \in POS \\ t_k \leq t}} PNews(k, t_k)\, e^{-\lambda(t-t_k)}, \quad t=1,\ldots,630 \qquad [3.4]$$

$$NImpact(t) = \sum_{\substack{k \in NEG \\ t_k \leq t}} NNews(k, t_k)\, e^{-\lambda(t-t_k)}, \quad t=1,\ldots,630 \qquad [3.5]$$

The arrival of more news items lead to higher values of accumulation; this therefore takes into account the news intensity, that is, the news flow. More importantly, the impact scores from previous news items are also included in the calculation of impact score at time *t*. Generally after 90 minutes the impact of a news story will have diminished to nothing. The impact is illustrated in Figure 3.7 which shows the impact score (with attenuation and accumulation) for the asset JP Morgan during the month of August 2008, with positive and negative sentiment represented separately.

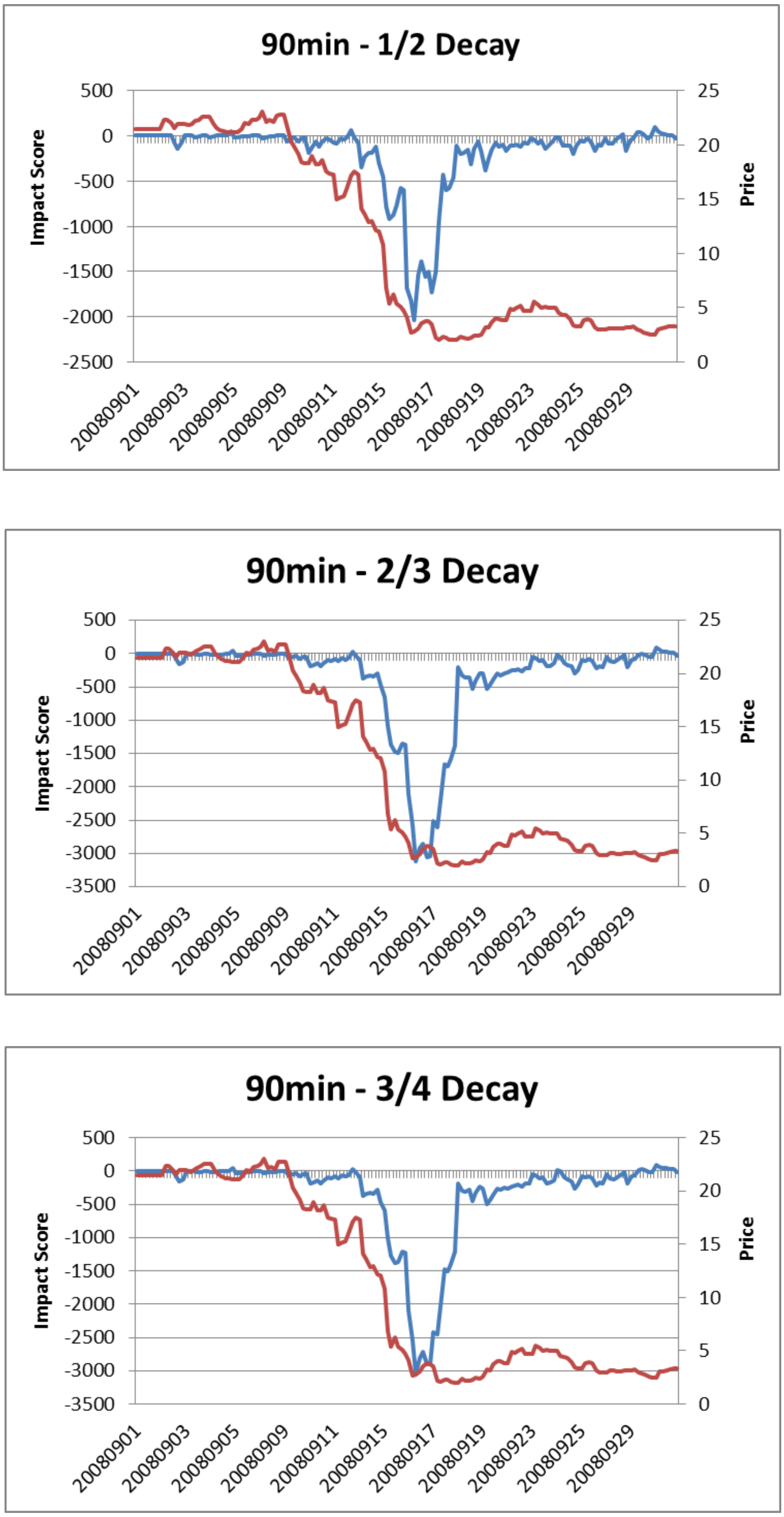


Figure 3.5: The decay of sentiment scores every 90 minutes for AIG over the month of September 2008 combined with different decay rates of a half, two-thirds and three-quarters respectively. The red line depicts the bid price.

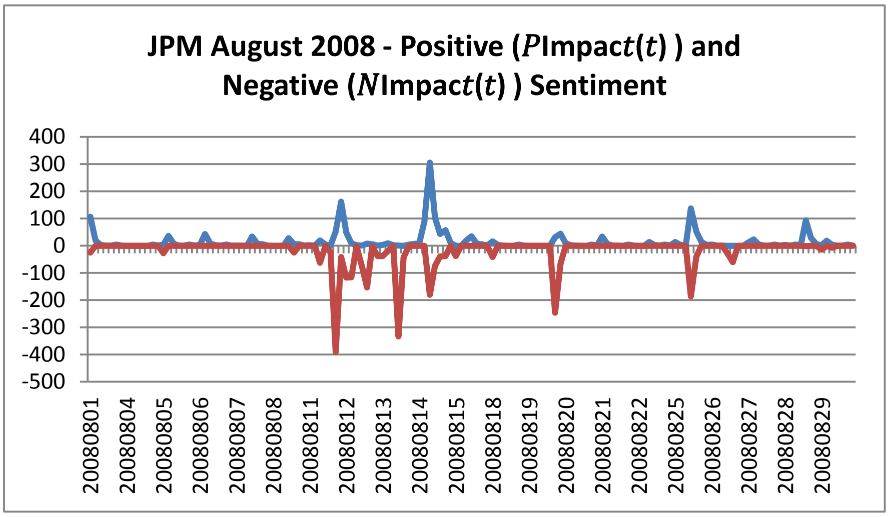


Figure 3.6: JP Morgan August 2008: News impact score (accumulated and aggregated) for positive (blue line) and negative (red line) sentiment respectively.

In order to compute the aggregated impact of all the news items which arrive during the day, we sum the positive and negative sentiments in each time bucket for the duration of the entire trading day. This gives us the impact score as set out in equation 3.6.

$$Impact_t = PImpact(t) + NImpact(t)\ , \quad t = 1, \ldots, 630 \qquad [3.6]$$

For convenience, we express *Impact(t)* also as $Impact_t$. Should there be no relevant news to appear throughout a day, then the impact score for the day equates to zero. Similarly, if in a time bucket no relevant news items are published, then $Impact_t$ is the total decayed value of previous news or zero if it is the first time bucket of the day.

Depending on the objectives of the researcher, it is also viable to consider the impact of news sentiment as two separate variables i.e. *PImpact(t)* and *NImpact(t)*, through logistic regression or the introduction of dummy variables. More noticeably, if the volume of news is a significant factor to be investigated then this is an appropriate approach.

In order to deduce whether causality exists between the impact score that we created and asset prices, once again a χ^2 test is performed. However, the setup of the experiment is simpler in this case as both prices and impact scores are presented in the same frequency – minute bar. Initial observation of the contingency table shows a strong relationship between the two random variables indicated by the largest numbers falling within the highlighted areas (see Table 3.2). The area highlighted in red constitutes the lowest prices with the lowest impact scores (values belonging to the smallest quintile). Conversely, the area highlighted in green represents the highest asset prices and the highest impact scores. That is to say low prices are reflected by low impact scores and high prices are reflected by high impact scores. Furthermore, this test is conducted between the asset price and lagged impact scores therefore, it can be perceived that such impact scores may possess predictability properties on asset prices. The yellow cross-section highlights the median values.

	I1	I2	I3	I4	I5
C1	6895	4934	24394	14227	267
C2	992	2013	5113	8092	0
C3	843	810	7440	15970	0
C4	835	1705	8747	16556	178
C5	1374	1911	9248	38285	84

Table 3.1: Contingency table for close price of AIG in 2008 and its respective impact score, split into quintiles. I1-I5 represent the impact score quintiles and C1-C5 represent the close price quintiles.

The null hypothesis for the χ^2 test is that the impact score and assets prices are independent. Testing at the 1% level with a test statistic of 27274.82 we reject the null hypothesis. Taking into consideration the observations made above, it is concluded that a casual effect between the factors of asset price and impact score could possibly exist. Therefore, with this evidence we proceed to investigate the relationship further and for subsequent chapters of this thesis all referral of sentiment is interpreted as the impact score.

Chapter 4

Univariate Predictive Model for Asset Behaviour

4.1 Introduction

In Chapter 1 we had given an overview of why sentiment analysis based on news has an important bearing on financial decision making. In this section we develop this theme further and introduce a univariate predictive model which connects news sentiment to asset dynamics. The univariate model uses return, volatility and liquidity to characterize asset behaviour. In this chapter, therefore, we adopt the impact score to construct the model which connects news sentiment with asset dynamics.

A major plank in the development and application of behavioural finance is the consideration of bounded rationality as introduced by Nobel laureate Herbert Simon (Simon, 1964). It follows from the theme of bounded rationality and later works of behavioural theorists/economists, Kahneman and Tverskey, Sheffrin, Shiller that human beings in general, and retail investors in particular, are influenced by various psychological imperatives of 'fear, greed and exuberance'. This is in sharp contrast with the postulates of neoclassical theories of rational behaviour and scrupulous application of logic in decision making. Thus behavioural finance in many ways determines the risk attitudes and the investment goals of (high net worth) individuals: so called HNIs who account for the majority of the invested wealth. Furthermore, this field also reinforces the importance of sentiment and investor psychology in market behaviour. Examples of such work include Shiller (2000) and Hais (2010) where the irrational contrarian and herd behaviour of investors are discussed.

Today the availability of sophisticated computer systems facilitating high frequency trading (Goodhart and O'Hara, 1997) as well as access to automated analysis of news feeds (Tetlock, 2007, Mitra and Mitra, 2011) set the backdrop for computer automated trading which is enhanced by news. How investment strategies may harness sentiment of news events and also that of the market continue to be actively studied and reported by researchers in the investment community (Peterson, 2007, Kahn, 2013, Dion, 2013, and Hafez, 2013). In this study we investigate how predictive models of asset behaviour can be used as a precursor to developing trading strategies.

Characterising Asset Behaviour

The majority of research on asset behaviour has focused on analysing historical market data and constructing models that best represent the information held in such datasets. The key features studied in these models are return and volatility of stock price, which are considered in the decision making stage of trading. In our study we introduce liquidity with a view to enhance these classical methods. Thus our model looks at three characteristics of an asset, namely return, volatility, and liquidity. The consideration of the additional parameter, liquidity, provides knowledge on the condition of markets and more importantly indicates whether a profitable signal can be successfully executed or not. Simply being able to determine a profitable position in the markets, through observation of stock price return and volatility, is not sufficient if in fact the actual trade is not available.

Asset price parameters have been traditionally represented by two methods: predictive modelling and explanatory modelling. One of the most well-known explanatory models is the factor model (Fama and French, 1992). They capture return by extending the Capital Asset Pricing Model (CAPM), which was independently proposed by Treynor (1961), Sharpe (1964), Lintner (1965), and Mossin (1966), to include market capitalisation size and book-to-market ratio as explanatory factors. Such types of models can be categorised under three groups, that is macroeconomic, fundamental and statistical factor models depending on the choice and nature of these factors and how the respective models are calibrated. Although factor models have dominated this field of finance, a weakness is in their failure to quickly update changes in market conditions. The structure of these models is only single period inhibiting the incorporation of relevant past information at a sufficient speed. Parameters are updated through calibration but only at a slow pace where the model adapts. Mitra, Mitra and diBartolomeo (2009) have shown how the incorporation of news enhances the results of factor models and leads to an early prediction of changes in volatility. We believe that news sentiment should also be considered as an explanatory factor in the pricing of assets and therefore return. However, our approach turns to another class of models that better predict volatility in a time-varying framework – generalised autoregressive conditional heteroskedasticity (GARCH) models.

Researchers in this domain (Ho, Shi and Zhang, 2013) have shown that by incorporating the additional factor of news sentiment in GARCH models, the near term volatility can be estimated using past news events over many periods. This near term estimation may be obtained using a number of alternative predictive models. Established models in this class are linear regression models (Stambaugh, 1999, Robertson and Wright, 2009), autoregressive (AR) models, moving average (MA) models and generalised autoregressive conditional heteroskedasticity (GARCH) (Bollerslev, 1986) models for volatility in particular. Our work utilises the family of AR and GARCH models within which news impact is introduced as an exogenous variable. We recall that news impact is a derived measure which takes into consideration (i) news sentiment and (ii) news flow, that is, volume of news (see section 3.4). The predictive models of asset behaviour are used as scenario generators. Scenarios are discrete realizations of an asset's characteristics (return, volatility and liquidity, in our case) which are used ex-ante for asset allocation in the face of uncertainty and ex-post for the simulation/evaluation of risk and other performance statistics.

Role of Liquidity in Asset Behaviour

It is important for traders to observe the liquidity of assets as it allows them to monitor the financial markets and assess the costs involved in a transaction. Moreover, the availability of liquidity is crucial to traders and brokers in their trading activities. Thus many recent studies have considered the implications of news on liquidity. For instance, Gross-Klaussman and Hautsch (2011) use the bid-ask spread, trading volume and market depth as proxies for liquidity with results showing greater increase of bid-ask spreads during news releases as opposed to market depth which does not differ much. Furthermore, Riordan et al. (2013) find that liquidity increases with news releases that have positive or neutral sentiment whereas negative news sentiment gives a corresponding decrease in liquidity.

The topic of alternative and relevant measures of liquidity is presented earlier in section 2.3; here we consider it again briefly to explain its context in the predictive model.

It is well accepted in the trading community that liquidity has the role of monitoring market conditions and assessing the viability of orders decided by trading algorithms by taking into consideration the spread and the depth of the market. Given that margin requirement is an important and defining aspect of trading, it can be argued that liquidity is an important determinant in deriving trading strategies and should be introduced as a parameter in predictive models. There are a variety of definitions which explain the role and measure of liquidity mainly in the context of trading. Spread measures view liquidity from the point of the cost that one has to bear for immediate trade, in other words the viability of orders, with typical measures being the effective spread and the bid-ask spread. Depth measures consider liquidity as the effect of large orders on a particular asset hence looking at market conditions, and are often measured using traded volume, order volume or Kyle's λ (Kyle, 1985).

Modelling Architecture and Choice of Assets

For the empirical study reported in this chapter we have considered 53 highly traded stocks taken from two exchanges, namely London Stock Exchange (FTSE) and New York Stock Exchange (Dow Jones); these stocks belong to nine industry sectors. Our aim is to find out whether taking into consideration news flow and news sentiment in addition to the market data leads to superior prediction of asset behaviour. We study the behaviour of equities as these are the most actively traded assets in the high frequency world and our aim is to ultimately derive models that can predict asset behaviour in high frequencies. This framework is highly applicable and relevant to the trading world as it supports decision-making models that is, strategies, which select trading portfolios with multiple assets.

4.2 Data

Our modelling architecture uses two streams of time series data: (i) The market data which is given at the minute bar level and includes bid price, ask price and the execution price, (ii) News metadata as supplied by Thomson Reuters. The structure and nature of the news metadata follows the description provided in section 3.1.

Market Data

The high frequency intraday market data is compiled on a minute-bar scale for 53 assets and covers nine sectors from banking, retail, oil to technology and communication. The data fields of the market data are set out in Table 4.1.

Data Field	Field Name	Description
1	#RIC	Reuters instrument code individually assigned to each company.
2	Date	In the format DD-MM-YYYY.
3	Time	In the format hh:mm:ss, given to the nearest minute.
4	GMT Offset	Difference from Greenwich Mean Time.
5	Type	Type of market data – in this case "Intraday 1 min"
6 - 9	Open; High; Low; Last	Open, high, low and last prices for the corresponding minute
10	Volume	Volume of trades in one minute
11	Ave. Price	Average price
12	VWAP	Volume weighted average price calculated for the corresponding minute
13	No. Trades	Number of trades in one minute
14 - 17	Open Bid; High Bid; Low Bid; Close Bid	Open, high, low and close bid prices for the corresponding minute
18	No. Bids	Number of bid orders placed
19 – 22	Open Ask; High Ask; Low Ask; Close Ask	Open, high, low and close ask prices for the corresponding minute
23	No. Asks	Number of ask orders placed

Table 4.1: Description of all the data fields for a company in the market data.

For those minutes that are missing in the data, the last price from the previous minute is used. Such situations occurred in the data as a reflection of no bid or ask prices during that time.

The Chosen Data set

A selection of 53 assets from FTSE100 and Dow Jones 30 across 9 different sectors is chosen for our empirical study. Table 4.2 lists these sectors and states the number of companies chosen within each sector. Summary statistics of the selected assets are given in Table 4.3. Companies with large market capitalisation were picked for their wide coverage of news; this consequently guarantees a sufficient number of data points in the time series of news metadata. In addition, the total market capitalisation of all selected assets is $4.36trillion (US stocks: $3.34trillion, UK stocks: $1.01trillion). This is a good representation of the entire stock market in the US and the UK as a large proportion of the market has been taken into account.

Sector	**Description**	**No. of companies chosen**
1	Banking	8
2	Insurance	6
3	Pharmaceuticals	6
4	Oil & Gas	7
5	Manufacturing	3
6	Retail	9
7	Telecommunications	4
8	IT & Technology	6
9	Media	4
		Total: 53

Table 4.2: Break down of the chosen assets by their respective sectors.

From the data fields described in Table 4.1, only close price, bid price and ask price are used. The exact period selected for model fitting is 2 January to 31 December 2008. From the minute bar data, we extract the prices from 08:00-18:30 each day to make up a **trading day** such that pre-trade and post-trade hours are included. The reason for this extension of hours is so that any news sentiment captured outside of trading hours can still be incorporated into our predictive models to avoid any loss of information (impact of news). Close prices are used to calculate log-return and volatility.

The general trend in asset prices for 2008 can be seen from Figure 4.1; since this was a period of negative (worsening) sentiment all prices are found moving downward. Simultaneously on a secondary Y-axis, the number of news for the same period is plotted showing the impact of news on asset prices.

The news metadata for the chosen assets were selected under the filter of relevance score, that is, any news item that had a relevance score under the value of 70 was ignored and not included in the data set. This ensured with a high degree of certainty that the sentiment scores to be used are indeed focused on the chosen asset and is not just a mention in the news for comparison purposes for example. In Table 4.3, it is observed from the column of relevant news that the number of news items for each asset is considerably reduced, even halved in some cases, once the filter of relevance is applied. Thus indicating that news sentiment of low relevance will definitely not be included in the chosen dataset.

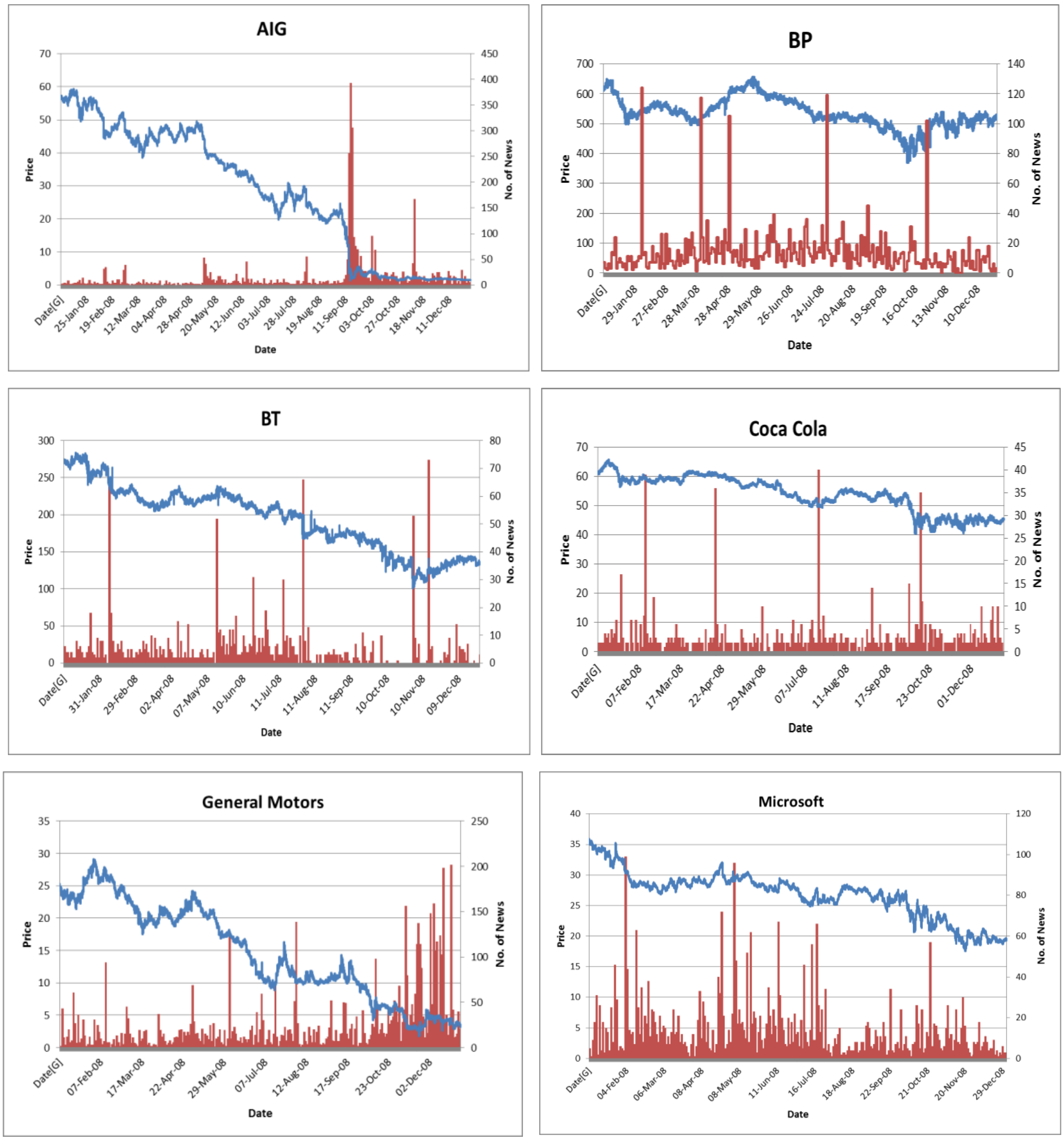


Figure 4.1: Closes prices (blue) plotted with number of news items (red) for AIG, BP, BT Group, Coca Cola, General Motors and Microsoft in the year 2008.

In Figure 4.2, we display the logarithmic return against the logarithmic number of news; a negative relationship is clearly evident. That is to say, a company with a large amount of news will have lower average annualised return than those with less news. It is not surprising to observe this effect for the year 2008, which is the sample used for testing; this was a period of global financial turmoil and so often news regarding a company would be portraying negative sentiment which is reflected in low returns.

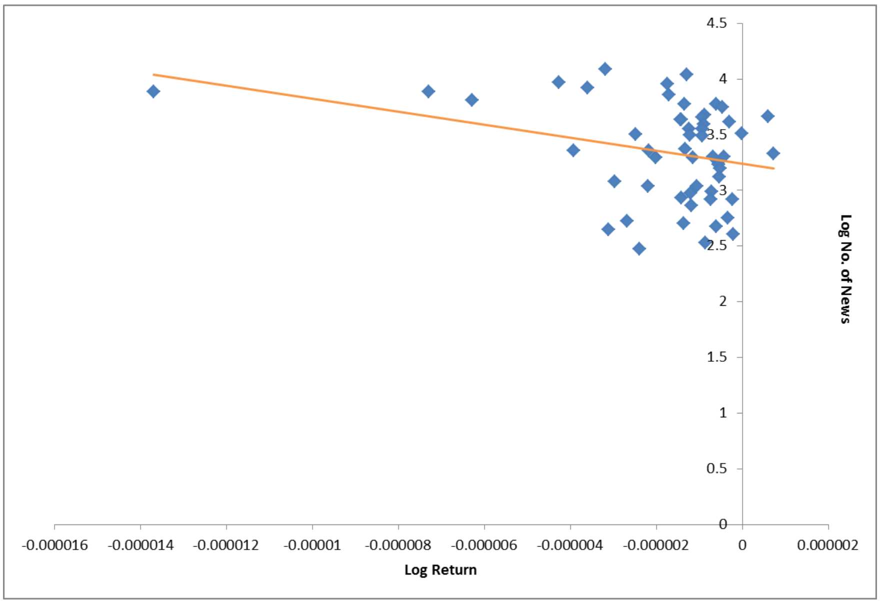


Figure 4.2: Graph showing the relationship between log return and log number of news.

Final preparation for the data is to align the frequency of sentiment scores to the trading hours of 08:00-18:30. Any news item released before or after these trading times were summed and bucketed in the next time period, which may be the following day. As a consequence, there is no discarding of news sentiment which could be influencing the price and return thus, not a single piece of news data is ignored.

Preliminary statistical tests are carried out on the chosen data set to determine if the asset price returns time series are stationary. To do this we implement the augmented Dickey-Fuller (ADF) test to see if a unit root exists in the autoregressive process. If it is present then the model is deemed non-stationary. The test is carried out under a hypothesis testing procedure with the ADF statistic being a negative number and the null hypothesis states that the coefficient of the first lag $\gamma = 0$. The ADF test is an expanded version of the Dickey-Fuller test developed in 1979 by statisticians David Dickey and Wayne Fuller, where only the AR(1) model is considered. The choice of adopting the ADF test is because the time series model we need to test is of a large size. Running ADF tests on the return series for all 53 assets, we obtain consistent negative values for the test statistic throughout (ranging from -44.205 to -64.567). This leads to the interpretation that all series are stationary because the null hypothesis is rejected at the 1% significance level.

Company Name	Sector	Market Cap. ($billions)	No. of News	Relevant News	Sentiment
AIG	Banking	81.36	7648	3870	-69.47
American Express	Banking	54.04	2261	1110	-9.90
AT&T	Telecommunication	207.08	4338	2430	16.53
Bank of America	Banking	475.94	9361	3283	-59.31
Chevron	Oil and Gas	161.51	3936	1519	-15.98
Coca Cola	Retail	262.55	1981	860	1.25
Disney	Media	54.1	2331	1080	3.76
Exxon Mobil	Oil and Gas	381.37	5900	2219	-21.80
General Electric	IT & Techonology	0.37	7680	3738	10.01
General Motors	Manufacturing	8.8	12176	6282	-102.32
Hewlett-Packard	IT & Techonology	84.40	3115	1536	4.24
Home Depot	Retail	44.73	1574	767	-6.79
IBM	IT & Techonology	118.89	3537	1911	20.77
Johnson & Johnson	Pharmaceuticals	177.42	1994	780	-3.51
JP Morgan	Banking	179.65	10971	4808	-5.45
Merck	Pharmaceuticals	139.22	3187	1535	-16.69
Microsoft	IT & Techonology	272.21	7245	3615	29.70
Pfizer	Pharmaceuticals	165.62	3091	1639	-5.47
Procter & Gamble	Retail	179.25	1995	887	3.48
Travelers	Insurance	18.11	972	612	-0.43
Verizon	Telecommunication	111.05	4536	3183	45.14
Wal-Mart	Retail	166.42	4633	1663	-13.85
Admiral Group	Insurance	3.86	474	344	1.30
ARM Holdings	IT & Techonology	1.97	1089	963	3.71
AstraZeneca	Pharmaceuticals	54.22	2135	1286	-6.27
Aviva	Insurance	22.01	5916	4800	22.47
Barclays	Banking	38.75	8987	5594	-47.25
BG Group	Oil and Gas	60.31	1700	1124	9.68
BP Group	Oil and Gas	161.80	5536	3436	-7.13
BskyB	Media	13.71	832	381	1.04
BT Group	Telecommunication	23.46	1957	1325	5.02
Burberry	Retail	2.73	525	410	0.23
GKN	Manufacturing	2.43	443	332	1.12
GlaxoSmithKleine	Pharmaceuticals	99.64	3259	1869	-0.53
ITV	Media	3.14	1092	678	-4.21
Kingfisher	Retail	4.819	827	545	-1.58
Legal & General	Insurance	10.15	4292	3623	6.01
Llodys Banking	Banking	28.70	8282	5476	29.00
Next	Retail	3.576	943	708	-5.33
Old Mutual	Insurance	8.064	1205	1057	2.59
Petrofac	Oil and Gas	3.025	503	418	2.24
Rolls Royce	Manufacturing	11.81	860	393	2.81
Royal Bank of Scotland	Banking	48.03	6409	2629	-1.82
Royal Dutch Shell	Oil and Gas	185.09	4137	2184	-15.55
RSA Insurance	Insurance	7.19	403	318	1.48
Sage Group	IT & Techonology	4.06	340	274	0.17
Shire	Pharmaceuticals	8.09	563	380	2.37
Standard Chartered	Banking	33.54	2289	1333	-0.10
Unilever	Retail	32.32	1317	1015	8.42
Vodafone	Telecommunication	122.37	4712	1788	15.71
Whitbread	Retail	3.141	730	625	5.72
Wood Group	Oil and Gas	3.109	295	223	0.76
WPP	Media	9.79	3572	2113	18.00

Table 4.3: Summary table of descriptive data for chosen assets within the period of 02/01/2008 – 31/12/2008.

4.3 The Predictive Model

The measures used in the model

We describe the bucket size, stock price return, volatility, liquidity and impact score below. With the raw market data that is provided to us, we are able to calculate asset behaviour measures using close prices, bid prices and ask prices.

Bucket Size

Bucket Size = 1 minute; Data Frequency = Minute Bar.

A bucket is equivalent to one period of time t and so lags are also at the same time frequencies.

The **trading day** starts at 08:00 hours and ends at 18:30 hours thus in a trading day the total number of buckets is 630.

Any news and sentiment retrieved overnight (between 18:31 – 07:59 the following day) is put in the first time bucket of the following day i.e. 08:00. The same applies to weekend news, which is aggregated over both days and considered in the first minute when trading is resumed. This method of categorisation reflects our belief that reactions to news are reflected in stock price movements hours (or even days) after release.

Return

The return measure that we use in the model is the log-return calculated by the following equation

$$Log(R_t) = Log(P_t) - Log(P_{t-1}) \qquad [4.1]$$

where P_t is the close price at time *t*. For convenience, $Log(R_t)$ is denoted by R_t throughout the rest of this thesis.

Volatility

The volatility measure used in the model is calculated as a rolling standard deviation of log returns for one **trading day** leading to 630 data points.

Liquidity

Liquidity is represented by the spread of the bid and ask prices, which measures the cost one has to bear for immediate trade. Equation 4.2 gives the expression for bid-ask spread.

$$Spread_t = \frac{p_t^A - p_t^B}{p_t^M} \qquad [4.2]$$

where p_t^A is the ask price at time *t*, p_t^B is the bid price at time *t*, and p_t^M is the mid-price between the bid and ask price at time *t*. Thereafter, the bid-ask spread will be denoted by S_t for convenience.

Impact Score

In our prediction models we use the impact score described in section 3.4 as the variable which measures (quantifies) the impact of news. A series of scores is calculated for each asset. A fundamental importance of this manipulation in sentiment score is that it produces a series of data accounting for every minute in a trading day. Therefore, this perfectly permits a fusion of market and news sentiment data to take place as both data series are matched in frequency. Here we reiterate the equation for impact score as:

$$Impact_t = PImpact(t) + NImpact(t)\ , \quad t = 1,\dots,630 \qquad [4.3]$$

where positive and negative impact are calculated separately using the following equations,

$$PImpact(t) = \sum_{\substack{k \in POS \\ t_k \le t}} PNews(k, t_k)\, e^{-\lambda(t-1)}, \quad t=1,\dots,630 \qquad [4.4]$$

$$NImpact(t) = \sum_{\substack{k \in NEG \\ t_k \le t}} NNews(k, t_k)\, e^{-\lambda(t-1)}, \quad t=1,\dots,630 \qquad [4.5]$$

and let $PNews(k, t_k)$ denote the sentiment value of the k^{th} positive news arriving at time bucket t_k and k belongs to the set of news with sentiment value $SENT$>0; $PNews(k, t_k) > 0$, and let $NNews(k, t_k)$ denote the sentiment value of the k^{th} negative news arriving at time bucket t_k and k belonging to the set of news with sentiment value $SENT$<0; $NNews(k, t_k) < 0$; $1 \leq t_k \leq 630$. For convenience, we express *Impact(t)* also as $Impact_t$.

The Predictive Model

Predictive analytical models for stock price return and volatility, such as the Autoregressive Conditional Heteroskedasticity (ARCH) model (Engle, 1982) and Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model (Bollerslev, 1986), are well understood and applied extensively. These models exploit the techniques of time series analysis (Chatfield, 2009, Harvey, 1990) and are able to account for the correlation between each of the individual points in the time series of stock price returns by fitting them to models which are of an autoregressive (AR) nature. We have adopted this approach in the construction of our models for predicting return, volatility, and liquidity. However, our method differs from typical ARCH/GARCH models in the sense that it deals with two time series: (i) the first time series is market data, (ii) the second time series is the news metadata. This approach can be seen as an extension of GARCH and AR models with the variable impact score being the innovative addition. The conditions of the markets and feasibility of trades is reflected through liquidity, which can be determined from two perspectives: (i) the spread of the market, evaluated by the bid-ask spread, and (ii) the depth of the market, computed by the total volume of bids and asks. We set out below a detailed description of the models.

Lag Selection Process

The decision of how many lags to incorporate in the two AR models for return and liquidity were based upon several selection methods. The two most common and basic methods for lag order selection are the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). Through observation of these values for our asset return and liquidity series, it is not possible to make any conclusions in the choice of lag orders. Thus, other options have to be explored. Nowadays, more advanced methods are available such as the Lasso variable selection technique (Tibshirani, 1996). Lasso is a shrinkage and selection method for linear regression and is beginning to find its way into financial applications (Mahler, 2009). The process minimises the sum of squared errors, subject to a bound, by a tuning parameter on the sum of the absolute values of the coefficients. It is exactly this tuning parameter that controls the amount of shrinkage applied to the estimates and in some case coefficients become equal to zero. Hence, we adopt this method to determine the number of lags to include in our autoregressive models. The results presented a best choice of lag two for the return model and lag three for the liquidity model.

Asset Return

Taking into consideration the inclusion of two time series, we construct an AR(2) predictive model for return with the enhancement of a news impact score.

Let, R_t be the log return of stock prices at time period *t*

R_{t-i} be the log return of stock prices by lag *i*

$Impact_{t-1}$ be the news impact score of the previous time interval

θ_i be the weighting coefficients to be estimated for lag *i*

e_t be the error term at time *t.*

Therefore, the log-return R_t is given by the expression,

$$R_t = \theta_0 + \theta_1 R_{t-1} + \theta_2 R_{t-2} + \theta_3 Impact_{t-1} + e_t \qquad [4.6]$$

Asset Volatility

The prediction model for volatility is an extended GARCH(1,1) model, as before with the addition of the impact score. The choice of implementing the GARCH(1,1) model is because we feel it is the best for characterising and modelling volatility.

Let, σ_t^2 be the volatility at time *t*

ϵ_{t-1}^2 be the lagged log-return residuals

σ_{t-1}^2 be the lagged volatility

$Impact_{t-1}$ be the impact score of the previous time interval

$\alpha_i, \beta_1, \omega_1$ be the weighting coefficients to be estimated

u_t be the error term at time *t.*

Therefore, the volatility σ_t^2 is given by the expression,

$$\sigma_t^2 = \alpha_o + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \omega_1 Impact_{t-1} + u_t \qquad [4.7]$$

Asset Liquidity

We measure the liquidity of the chosen asset by the bid-ask spread and construct an AR(3) model with the addition of the impact score.

Let, S_t be the bid-ask spread at time *t*
S_{t-i} be the bid-ask spreads at lag *i*
$Impact_{t-1}$ be the impact score of the previous time interval
γ_i be the weighting coefficients to be estimated for lag *i*
η_t be the error term at time *t.*

Therefore, the model for S_t is as follows:

$$S_t = \gamma_o + \gamma_1 S_{t-1} + \gamma_2 S_{t-2} + \gamma_3 S_{t-3} + \gamma_4 Impact_{t-1} + \eta_t \qquad [4.8]$$

4.4 Computational Results and Validation

In Sample Results

The estimates for each weighting coefficient of the return, volatility and liquidity predictive models for the univariate case are presented in Figure 4.3, Figure 4.4 and Figure 4.5 respectively for all 53 assets. The estimates are displayed according to the colour scheme detailed in Table 4.5. From the fitting of the models it can be seen that news sentiment does indeed have an effect on the prediction of return, volatility and liquidity as all coefficient estimates are non-zero.

Model	**Parameter**	**Colour**
Return	θ_0 – Intercept θ_1 – 1st lag of return θ_2 – 2nd lag of return θ_3 – 1st lag of Impact score	Dark Blue Green Red Purple
Volatility	α_o- Intercept α_1 – 1st lag of return residuals β_1 – 1st lag of volatility ω_1 – 1st lag of Impact score	Orange Gold Turquoise Violet
Liquidity	γ_o- Intercept γ_1 – 1st lag of spread γ_2 – 2nd lag of spread γ_3- 3rd lag of spread γ_4 – 1st lag of Impact score	Light Blue Pink Yellow Light Green Grey

Table 4.4: Colour scheme applied in Figures 4.3, 4.4 and 4.5 to represent the coefficient estimates.

Furthermore, our statistical hypothesis tests highlight the significance of the news variable in the predictive models. In particular, the following test was set up to investigate the significance of news effect on the volatility model:

$$\mathrm{H}_0: \omega_1 = 0 \text{ vs } H_1: \omega_1 \neq 0 .$$

This deemed to be highly significant at the 5% significance level for all 53 assets (p-value = <$2\,e^{-16}$) and so the null hypothesis is rejected, concluding that the estimator for the coefficient of impact score is not equal to zero and therefore has an influence on the prediction of volatility.

All intercepts of return are estimated to be negative apart from a handful of assets. The results of Microsoft showed to be the most extreme with values far greater than the average estimate of -1.75e^{-6}**.** In fact, the performance of Microsoft consistently stood out in the regressions of all three measures.

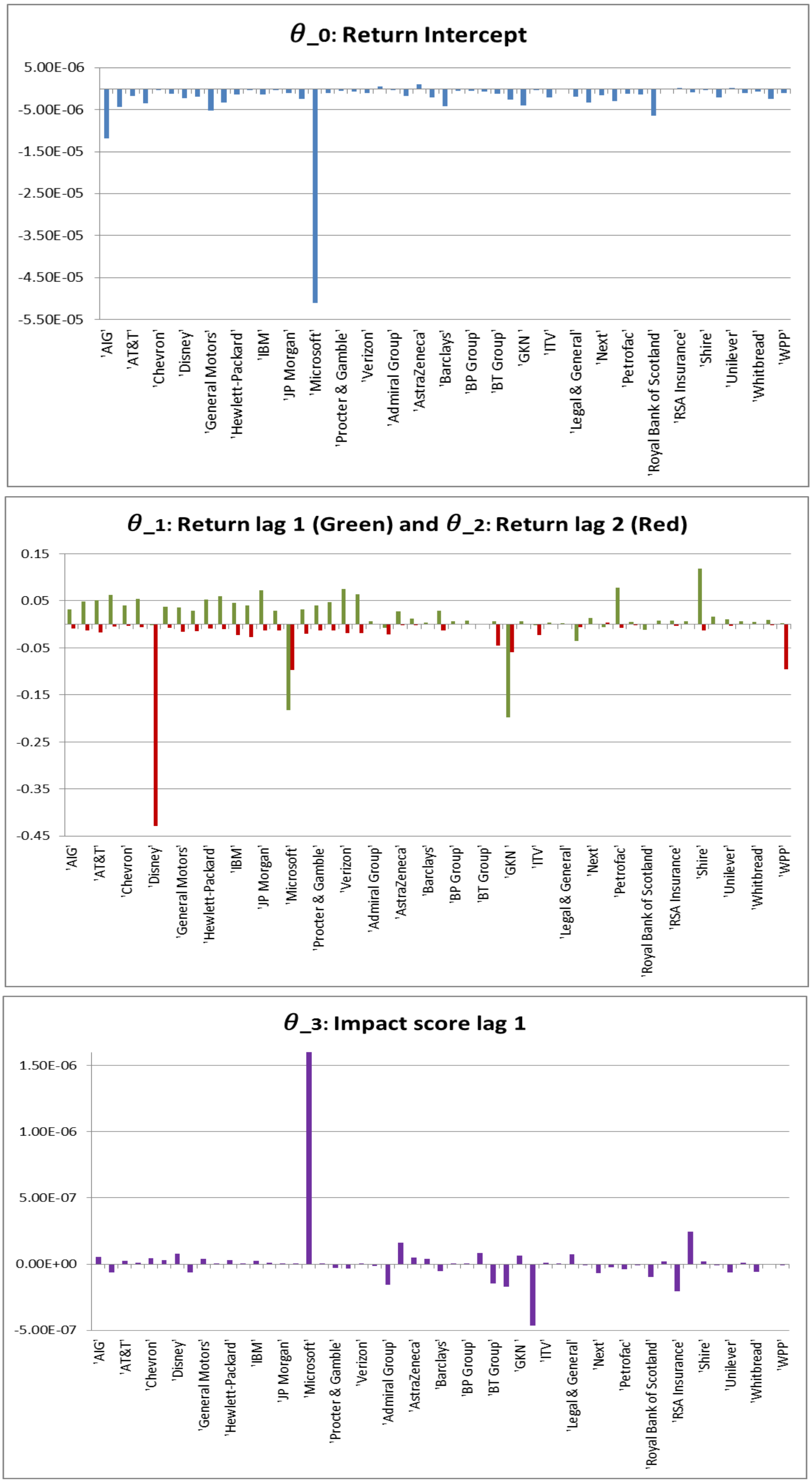


Figure 4.3: Distribution of 53 assets' estimated coefficients for the lagged returns and impact score in the news enhanced return model.

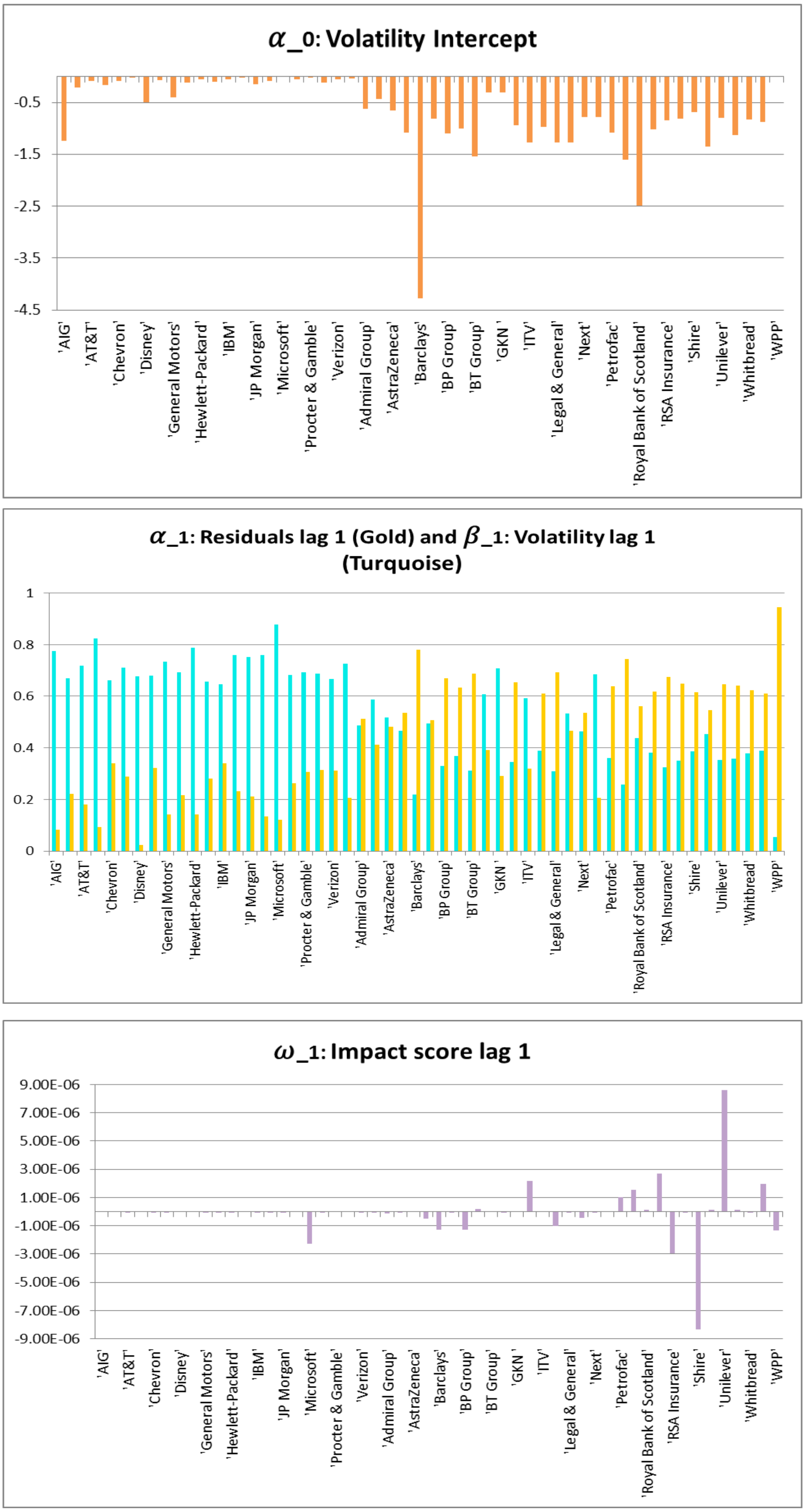


Figure 4.4: Distribution of 53 assets' estimated coefficients for the GARCH and impact score in the news enhanced volatility model.

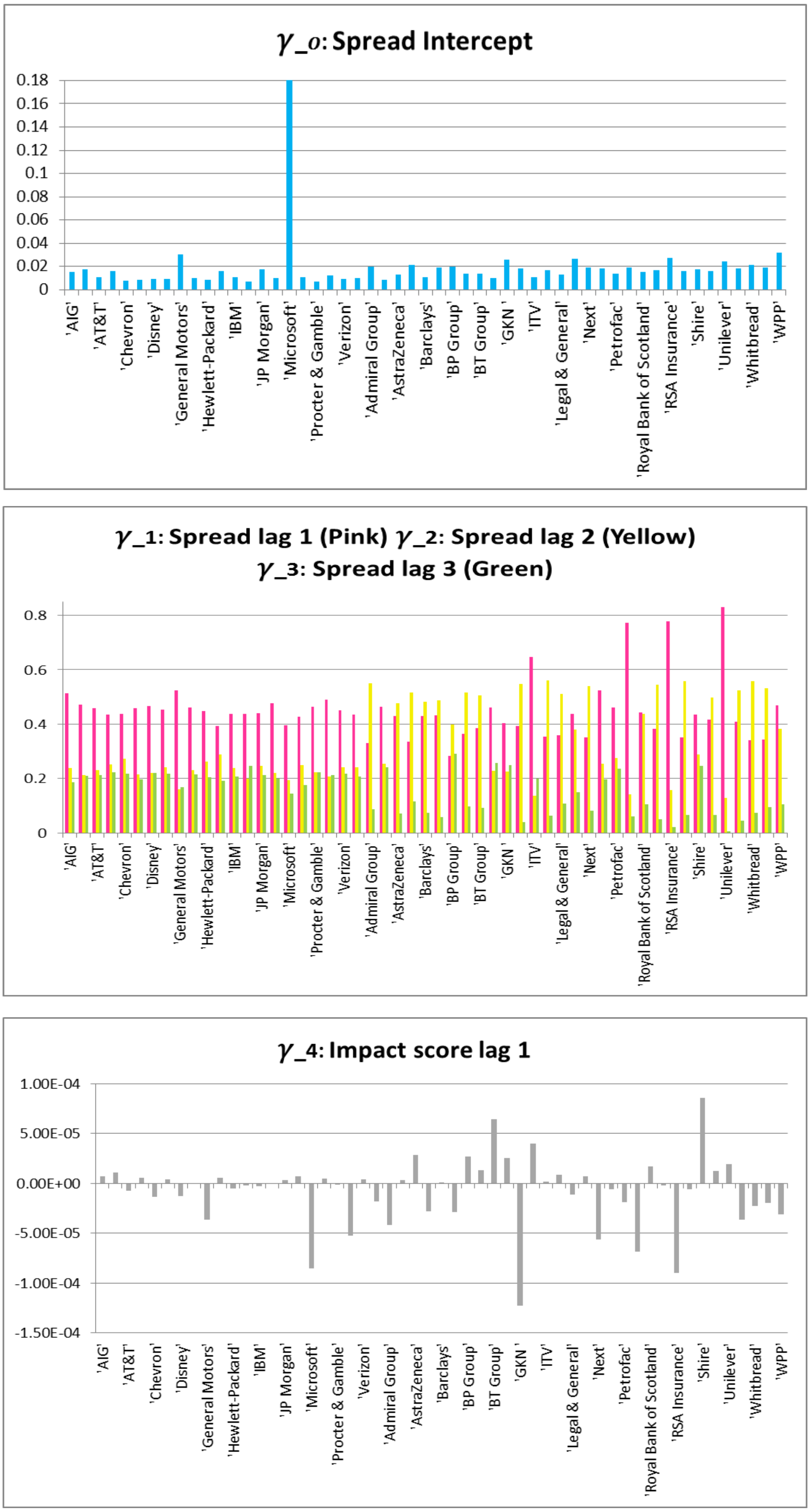


Figure 4.5: Distribution of 53 assets' estimated coefficients for the lagged spread and impact score variables in the news enhanced spread model.

Out-of-Sample Results

To understand the extent to which news impact enhances the results of these models, we use the fitted model to predict out-of-sample values. To analyse the performance of these news enhanced models, we take the benchmark to be predictive models that only consider market data (referred to as a market only model from here onwards). Specifically, the models described in section 4.3 with the exogenous variable of the impact score removed, which would simply be AR and GARCH models. Direct comparison is made between the two sets of models to discover whether the news enhanced models perform better and hence whether news impact has a positive influence on the prediction of return, liquidity and volatility. The out-of-sample period is taken as January – March 2009. The conclusions interpreted are that news enhanced predictive models do better predict future values of liquidity and volatility but not for stock returns. Furthermore, the inclusion of news sentiment to a predictive model, in the form of an impact score, does increase accuracy of predictions of liquidity and volatility.

The predictive power of the fitted models is represented as the difference between the residuals from the market only model and the market with news model, i.e. prediction error. Prediction performance of the return model produces inconclusive results as there is a variation of positive and negative residuals across the assets. The prediction in the direction of return, and therefore price, is accurate however the magnitude of the return can be erroneous. This applies to both the market only model and the news enhanced model with little difference between them, as can be seen from Figure 4.6 where errors range from $6e^{-6}$ to -$6e^{-6}$. However, there are instances where the added news sentiment impacts the prediction of return in a more positive manner than the benchmark model, for example the Banking sector (see Figure 4.9). Figure 4.6 plots the average difference in prediction errors for each asset from the chosen universe of assets.

Company	*Average Return Prediction Error of News Enhanced Model*	*Average Return Prediction Error of Market only model*
AIG	-1.7067e^{-5}	-1.3280e^{-5}
AT&T	-1.8239e^{-6}	-1.3818e^{-6}
American Express	-3.6277e^{-6}	-2.7099e^{-6}
Bank of America	-5.0896e^{-6}	-3.9985e^{-6}
Chevron	-1.2698e^{-6}	-9.2752e^{-7}
Coca Cola	-1.3013e^{-6}	-1.1237e^{-6}
Disney	-2.8490e^{-6}	-2.1118e^{-6}
Exxon Mobil	-9.5528e^{-7}	-6.4897e^{-7}
General Motors	-1.2052e^{-5}	-7.2911e^{-6}
General Electric	-3.4138e^{-6}	-3.16011e^{-6}
Hewlett-Packard	-1.4318e^{-6}	-1.1466e^{-6}
The Home Depot	-3.4235e^{-7}	-3.5743e^{-7}
IBM	-1.2500e^{-6}	-7.6193e^{-7}
Johnson & Johnson	-3.8503e^{-7}	-4.0296e^{-7}
JP Morgan	-1.3918e^{-6}	-1.3114e^{-6}
Merck	-2.3682e^{-6}	-2.4504e^{-6}
Microsoft	-8.8786e^{-5}	-1.9044e^{-6}
Pfzier	-9.6822e^{-7}	-9.8333e^{-7}
Procter & Gamble	-5.1619e^{-7}	-6.6652e^{-7}
Travelers	-6.0209e^{-7}	-6.6273e^{-7}
Verizon	-9.5462e^{-7}	-9.3701e^{-7}
Wal-Mart	8.8896e^{-7}	5.6024e^{-7}
Admiral Group	-1.1982e^{-6}	-5.9244e^{-7}
ARM Holdings	-9.6706e^{-7}	-1.0896e^{-6}
AstraZeneca	8.0655e^{-7}	7.1590e^{-7}
Aviva	-2.2480e^{-6}	-1.3181e^{-6}
Barclays	-4.7188e^{-7}	-1.6756e^{-6}
BG Group	-4.6441e^{-7}	-4.5952e^{-7}
BP Group	-5.1270e^{-7}	-5.0008e^{-7}
BskyB	-1.1601e^{-6}	-6.7193e^{-7}
BT Group	3.3406e^{-7}	-1.9840e^{-6}
Burberry	-2.3389e^{-6}	-2.8959e^{-6}
GKN	-4.0280e^{-6}	-3.7416e^{-6}
GlaxoSmithKline	-2.1747e^{-6}	2.4036e^{-8}
ITV	-2.0867e^{-6}	-2.0659e^{-6}
Kingfisher	-6.8699e^{-8}	-7.9559e^{-8}
Legal & General	-1.3537e^{-6}	-1.4074e^{-6}
Llodys Banking	-2.8758e^{-6}	-3.4989e^{-6}
Next	-1.2381e^{-6}	-1.0460e^{-6}
Old Mutual	-3.0012e^{-6}	-2.9140e^{-6}
Petrofac	-1.0774e^{-6}	-9.5803e^{-7}
Rolls Royce	-1.4019e^{-6}	-1.3850e^{-6}
Royal Bank of Scotland	-2.0462e^{-7}	-6.2298e^{-6}
Royal Dutch Shell	-2.7725e^{-7}	-3.2273e^{-7}
RSA Insurance	-2.8436e^{-7}	-1.5840e^{-7}
Sage Group	-6.3643e^{-7}	-8.1228e^{-7}
Shire	-4.4896e^{-7}	-5.1144e^{-7}
Standard Chartered	-2.0220e^{-6}	-2.0871e^{-6}
Unilever	-3.7355e^{-7}	-5.2755e^{-7}
Vodafone	-7.4097e^{-7}	-8.0920e^{-7}
Whitbread	-6.5725e^{-7}	-1.0460e^{-6}
Wood Group	-2.3476e^{-6}	-2.3480e^{-6}
WPP	-1.1461e^{-6}	-1.2901e^{-6}

Table 4.5: Average prediction errors of return for all 53 assets from the news enhanced predictive model and the market data only model.

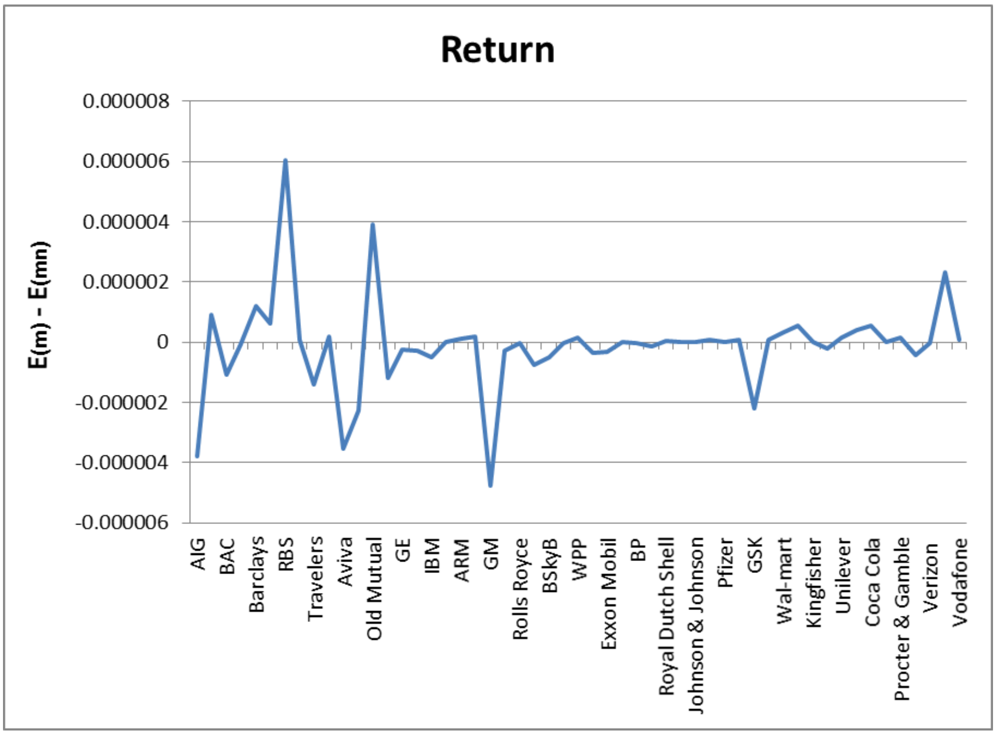


Figure 4.6: Average difference in prediction error between market only (E(m)) and news enhanced models (E(mn)) for log-return for all assets in the period January-March 2009.

For the case of volatility, all assets performed better in the news enhanced GARCH model as indicated by a positive difference in prediction errors throughout (see Figure 4.7). In other words, the residuals produced from the market only model were far greater than those from the news enhanced model, almost consistently by a factor of ten. Moreover, the plain market only model has a tendency to over predict the values of volatility. Therefore, the inclusion of news sentiment as an exogenous variable to the univariate GARCH model significantly reduces prediction errors and produces estimates far closer to the true values of volatility. This is illustrated in Figure 4.7 which plots the difference in prediction errors for all assets.

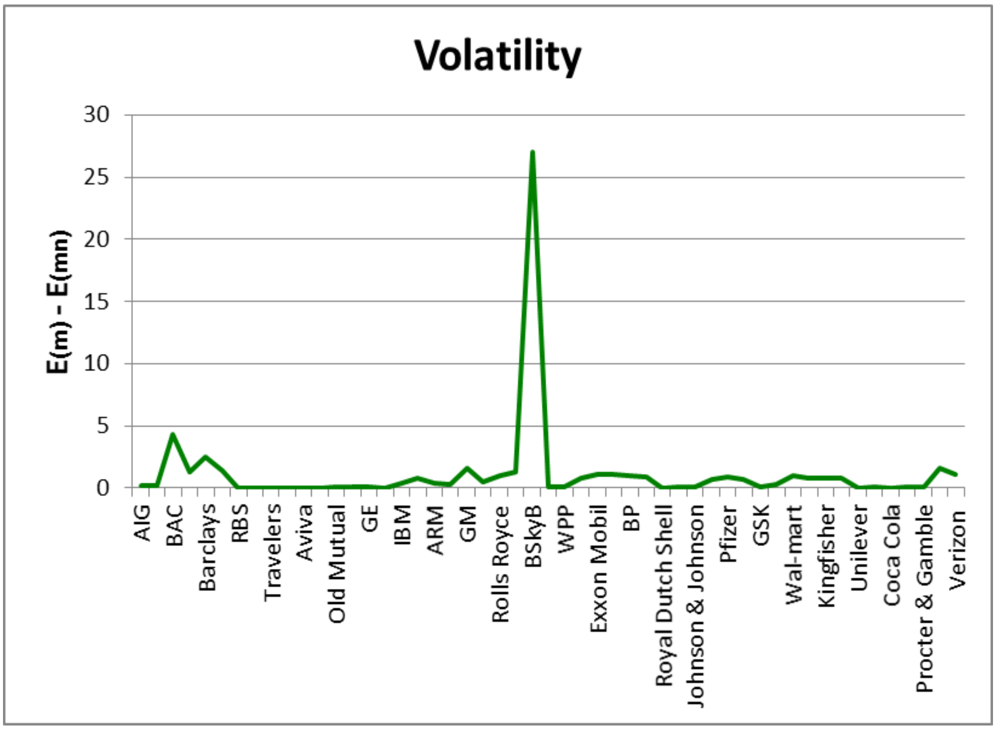


Figure 4.7: Average difference in prediction error between market only (E(m)) and news enhanced models (E(mn)) for volatility for all assets in the period January-March 2009.

Company	*Average Volatility Prediction Error of News Enhanced Model*	*Average Volatility Prediction Error of Market only model*
AIG	$6.4036e^{-6}$	1.2457
AT&T	$2.7483e^{-7}$	0.0903
American Express	$1.4115e^{-6}$	0.2071
Bank of America	$5.9190e^{-6}$	0.1685
Chevron	$2.3695e^{-7}$	0.0797
Coca Cola	$1.2501e^{-7}$	0.0242
Disney	$5.9238e^{-7}$	0.4963
Exxon Mobil	$1.8712e^{-7}$	0.0693
General Motors	$4.5241e^{-6}$	0.4072
General Electric	$1.0043e^{-6}$	0.1112
Hewlett-Packard	$3.2316e^{-7}$	0.0582
The Home Depot	$3.4888e^{-7}$	0.0969
IBM	$1.9280e^{-7}$	0.0516
Johnson & Johnson	$1.1121e^{-7}$	0.0190
JP Morgan	0.0606	0.2024
Merck	$4.1994e^{-7}$	0.0836
Microsoft	0.0148	117.0103
Pfzier	$2.6102e^{-7}$	0.0533
Procter & Gamble	0.0215	0.0215
Travelers	$5.5253e^{-7}$	0.1190
Verizon	$2.0731e^{-7}$	0.0550
Wal-Mart	$2.0420e^{-7}$	0.0323
Admiral Group	$2.2579e^{-6}$	0.6276
ARM Holdings	$2.7135e^{-6}$	0.4298
AstraZeneca	$3.7780e^{-5}$	0.6536
Aviva	$6.1802e^{-5}$	1.0757
Barclays	0.0003	4.2774
BG Group	$4.3482e^{-6}$	0.8180
BP Group	$1.2957e^{-5}$	1.0923
BskyB	$1.8036e^{-6}$	0.9958
BT Group	$8.2825e^{-5}$	1.5400
Burberry	$3.0501e^{-6}$	0.3065
GKN	$3.9367e^{-6}$	0.3130
GlaxoSmithKline	$1.3452e^{-5}$	0.9310
ITV	$4.7490e^{-6}$	1.2772
Kingfisher	$-1.2436e^{-6}$	0.9679
Legal & General	$7.4016e^{-6}$	1.2680
Llodys Banking	0.0004	1.2641
Next	$4.5398e^{-5}$	0.7806
Old Mutual	$3.8272e^{-6}$	0.7821
Petrofac	$6.4614e^{-6}$	1.0737
Rolls Royce	$5.7371e^{-6}$	1.6029
Royal Bank of Scotland	0.0009	2.4976
Royal Dutch Shell	$9.5086e^{-6}$	1.0253
RSA Insurance	$1.6118e^{-6}$	0.8367
Sage Group	$2.8322e^{-6}$	0.8084
Shire	$-2.4374e^{-5}$	0.6839
Standard Chartered	$2.7873e^{-6}$	1.34446
Unilever	$-1.7670e^{-5}$	0.7890
Vodafone	$5.8874e^{-6}$	1.1222
Whitbread	$2.7631e^{-6}$	0.8245
Wood Group	$1.6346e^{-5}$	0.8695
WPP	$2.3758e^{-5}$	27.0141

Table 4.6: Average prediction errors of liquidity (bid-ask spread) for all 53 assets from the news enhanced predictive model and the market data only model.

The difference in errors for liquidity (bid-ask spread) is relatively smaller than those seen for volatility, with the majority of values falling within 3 decimal points (see Figure 4.8). It can be observed that positive values dominate in the plot of errors between the two models and an overestimation by the market only model. Negative values do appear but only for major falls in spread. One explanation is that periods of low spread indicate a high volume of trading activity and such active trading periods are no longer instigated purely by the release of news. Another influential factor could be bearish market conditions which cannot be accounted for in a news enhanced model.

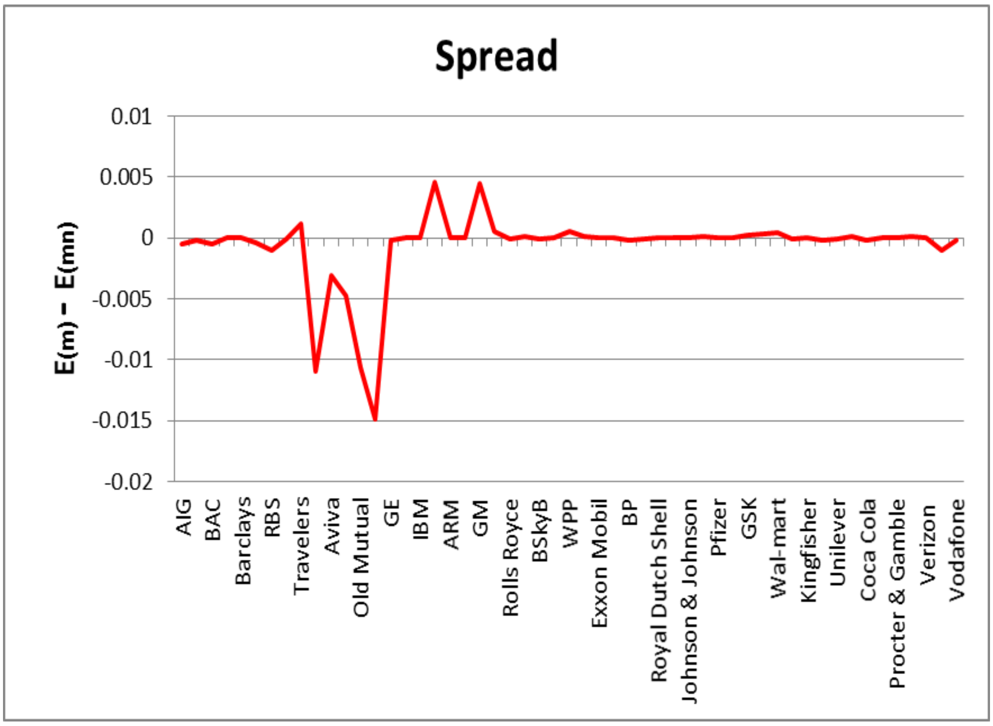


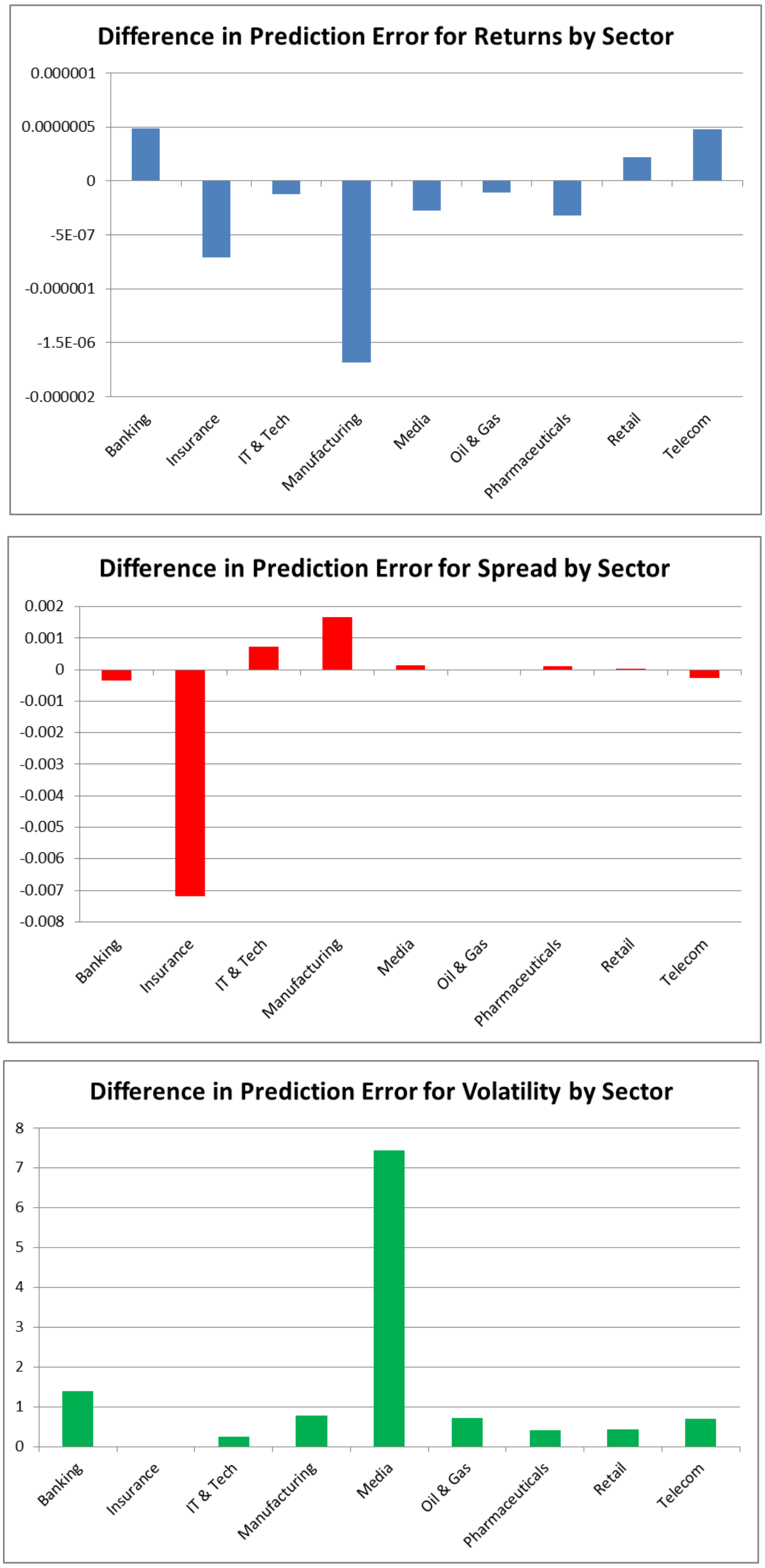
Figure 4.8: Average difference in prediction error between market only (E(m)) and news enhanced models (E(mn)) for bid-ask spread for all assets in the period January-March 2009.

Company	*Average Spread Prediction Error of News Enhanced Model*	*Average Spread Prediction Error of Market only model*
AIG	0.1618	0.1622
AT&T	0.0872	0.0892
American Express	0.1914	0.1915
Bank of America	0.2742	0.2747
Chevron	0.0970	0.0990
Coca Cola	0.0649	0.0665
Disney	0.1040	0.1063
Exxon Mobil	0.0822	0.0841
General Motors	0.1897	0.1879
General Electric	0.1469	0.1504
Hewlett-Packard	0.0935	0.0956
The Home Depot	0.1125	0.1152
IBM	0.0822	0.0842
Johnson & Johnson	0.0610	0.0624
Merck	0.0938	0.0960
Microsoft	-0.9557	-0.8836
Pfzier	0.0749	0.0768
Procter & Gamble	0.0727	0.0727
Travelers	0.1401	0.1432
Verizon	0.0823	0.0847
Wal-Mart	0.0735	0.0747
Admiral Group	0.2560	0.2630
ARM Holdings	0.1371	0.1412
AstraZeneca	0.4862	0.4957
Aviva	0.4805	0.4898
Barclays	1.3435	1.3435
BG Group	0.5347	0.5461
BP Group	0.5080	0.5183
BskyB	0.4563	0.4655
BT Group	0.6583	0.6714
Burberry	0.2144	0.2191
GKN	0.1683	0.1723
GlaxoSmithKline	0.4742	0.4841
ITV	0.1536	0.1581
Kingfisher	0.5065	0.5169
Legal & General	0.3794	0.3884
Llodys Banking	0.7773	0.7769
Next	0.6263	0.6384
Old Mutual	0.2690	0.2759
Petrofac	0.1669	0.1728
Rolls Royce	0.4637	0.4749
Royal Bank of Scotland	0.7064	0.7053
Royal Dutch Shell	0.5745	0.5854
RSA Insurance	0.2817	0.2893
Sage Group	0.3838	0.3923
Shire	0.2885	0.2959
Standard Chartered	0.3456	0.3455
Unilever	0.3965	0.4063
Vodafone	0.6103	0.6235
Whitbread	0.3881	0.3963
Wood Group	0.2358	0.2420
WPP	0.3722	0.3811

Table 4.7: Average prediction errors of all 53 assets from the news enhanced predictive model and the market data only model.

Figure 4.9 shows the performance of predictions for log-return, liquidity and volatility by sector. Each value is the average difference in prediction error between the market only model and the news enhanced model across all assets within that sector. Not only does it clarify the superiority of prediction power by the news enhanced model for liquidity and volatility, but it also shows the best performing sectors within each measure. Banking and telecommunications ranked top for return predictions and manufacturing was worst. By contrast, the best performing sector for liquidity predictions was manufacturing as it had the highest positive error with a value of 0.0017 compared with insurance that had the most negative value of -0.0072. All sectors returned a positive error for volatility predictions, but the media industry distinctively outperformed the rest with an average error of 7.45, which is wildly distant from the true values of volatility that lie between 0-0.01.

Statistical measures such as root mean squared error (RMSE) and mean absolute error (MAE) were also calculated to compare prediction accuracy between the models. The results exhibit a very strong indication of superior performance in the prediction of volatility with the news enhanced model. Table 4.5 lists the rankings of prediction performance for each asset according to return, volatility and liquidity. Through ranking the forecasting performance of each asset with respect to their RMSE, an obvious pattern emerges between return and volatility prediction performances. From Table 4.5 it can be seen that there is a reversal in rankings for return and volatility, particularly in the highest and lowest ranks, with Microsoft showing the largest contrast in performance ranking last place for return prediction but first place for volatility prediction.



Difference in Prediction Error for Returns by Sector

Difference in Prediction Error for Spread by Sector

Difference in Prediction Error for Volatility by Sector

Figure 4.9: Average differences between out-of-sample prediction errors for market only and news enhanced model for log-return, bid-ask spread and volatility during the period January-March 2009.

Asset	Return Prediction Rank	Liquidity Prediction Rank	Volatility Prediction Rank
American Express	1	39	36
Aviva	2	29	15
Disney	3	9	31
Chevron	4	42	44
Hewlett-Packard	5	25	46
JP Morgan	6	53	39
Legal & General	7	11	9
General Electric	8	41	40
ARM Holdings	9	15	32
Bank of America	10	49	37
Burberry	11	46	35
The Home Depot	12	24	41
Travelers	13	36	38
Whitbread	14	38	22
ITV	15	14	8
Verizon	16	12	47
BT Group	17	45	6
Vodafone	18	1	12
Rolls Royce	19	30	5
Sage Group	20	21	24
Pfzier	21	26	48
BP Group	22	31	13
Royal Dutch Shell	23	22	16
Wood Group	24	20	20
Next	25	28	27
Kingfisher	26	37	18
BG Group	27	32	23
Standard Chartered	28	5	7
WPP	29	40	2
Coca Cola	30	8	51
GKN	31	2	34
Johnson & Johnson	32	23	53
Merck	33	10	43
Petrofac	34	13	14
Llodys Banking	35	4	10
AstraZeneca	36	3	29
AT&T	37	33	42
Old Mutual	38	17	26
Wal-Mart	39	51	50
Shire	40	48	28
Admiral Group	41	34	30
BSkyB	42	7	17
IBM	43	19	49
Procter & Gamble	44	16	52
Unilever	45	35	25
RSA Insurance	46	47	21
General Motors	47	52	33
Royal Bank of Scotland	48	43	4
Barclays	49	27	3
Exxon Mobil	50	18	45
AIG	51	44	11
GlaxoSmithKline	52	6	19
Microsoft	53	50	1

Table 4.8: Forecasting performance ranking for each asset based on the root mean squared error for return, spread and volatility. Smallest errors receive the highest rank.

4.5 Summary

Behavioural models and sentiment analysis are gaining momentum and acceptance within the investment community. Our study sets out to consider news sentiment and its impact on asset behaviour. We have introduced a novel concept of **news impact** which takes into consideration (i) the volume of news and (ii) the decay of the effect of news sentiment over time. We are interested in the predictive analysis of asset behaviour: return, volatility and liquidity, so that this can be applied to construct trading strategies in an intraday setting. The findings are as follows. Enhancement of predictions of liquidity and volatility by inclusion of news metadata is found with supporting arguments for the representative data set. In particular the improvement of volatility predictions is substantial; this is in line with earlier studies as well as reported results by other investigators. Liquidity prediction is also improved but not as much as that of volatility. Price prediction does not show any improvement at all and this is commensurate with results reported by other researchers. One explanation could be that market participants react quickly which leads to revision of price in response to news; however, volatility and liquidity lack this behaviour and hence are better predicted by the incorporation of news.

Chapter 5

Multivariate Predictive Model using Bayesian Inference

5.1 Introduction

In this chapter we describe our empirical study of a (news enhanced) multivariate predictive model. We first discuss the context and relevance of the research problem and the study. From the perspective of designing a trading strategy it is necessary to predict the behaviour of a single asset or a small collection of assets which are used in the trading strategy. So for the collection of assets which defines the universe of the trading portfolio it becomes necessary to describe/predict the discrete realisations of these assets in a multivariate setting. These discrete realisations are also popularly known as scenarios and the corresponding model a scenario generator. Thus the multivariate model can generate scenarios which find use in asset allocation that is, trading strategies. Equally the scenarios are used in the simulation (computation) of the risk of a given trade/trading strategy. This sets out the context and the focus of the multivariate predictive model. In our predictive model, we follow the approach of the Black-Litterman model (1991) and combine the news sentiment data series with the estimated series of return, volatility and liquidity to create a new estimation in the form of a posterior distribution based on a mix of data. The models constructed are multivariate taking into account the prediction and correlation of several assets simultaneously.

Trading in the financial markets equates to rebalancing one or multiple (a few) stocks at a high frequency. The trading decision taken for each individual asset also needs to be coordinated with its counterparts dependent on their correlation. Therefore, the requirement of multivariate modelling arises. We fulfil this requirement of multivariate modelling by introducing three different models: multivariate linear models, multivariate GARCH models and vector autoregressive (VAR) models for asset return, volatility and liquidity respectively. The choice of each category of model for each of the three variables is justified in section 5.2

The Black-Litterman model estimates expected excess returns and covariance, which is applied as an input to an optimizer. Fisher Black and Robert Litterman claim to overcome three essential problems of portfolio construction. They are unintuitive, highly-concentrated portfolios, input-sensitivity and estimation error maximization. Beside these factors, the model also clearly specifies a method of incorporating investors' views into the return estimation model. The prior is chosen to be the CAPM equilibrium market portfolio. The Black-Litterman model enables investors to combine their unique views regarding the performance of various assets with the market equilibrium in a manner that results in intuitive, diversified portfolios. The same concept and argument are used in Bayesian Inference where additional information such as investors' views, expert opinions and industry experience are incorporated into the modelling process. In our work, we apply the idea of Black-Litterman (1991) and perform news sentiment based market data analysis for return, volatility and liquidity (spread) by using a formal Bayesian approach.

As we restated earlier, the cliché is "news moves the markets". Thus a news item affects (influences) multiple stocks which are under consideration. One of these effects is known as information spill over effect. This describes the instance when news for one asset has partial implications for another asset (non-announcing stock) which in turn affects the pricing of both stocks. Subsequently, the covariance between returns of the announcing stocks and market return increases.

Bayesian inference

Classical estimation techniques for fitting regression models typically test regression coefficients as unknown but fixed, and then proceed to implement frequents approach which assumes sufficient measurements to gather meaningful information about the unknown parameter. The least squares estimation method is an example of such an approach. Bayesian techniques however, treat the regression parameters as random variables and select priors according to the mean/median/mode of the distribution. An advantage of Bayesian inference is the ability to calculate confidence intervals in a straightforward manner. Bayesian predictive distributions are straightforward to calculate and summarize the fund manager or the investor's views of future return distributions (Jacquier and Polson, 2010, Barberis, 2000).

In the Bayesian approach, data is supplemented with additional information in the form of a prior probability distribution. According to Bayes’ theorem, the prior distribution, $p(\theta)$, for the parameters θ combined with the data's likelihood function, $p(data|\theta)$, yields the posterior distribution about the parameters, $p(\theta|data)$. Mathematically, Bayes' theorem states that the posterior distribution is proportional to the product of the prior distribution of parameters and the likelihood function from data:

$$p(\theta|data) \propto p(\theta) * p(data|\theta)$$

The prior can take different functional forms depending on the domain and the information that is available a priori. But a conjugate prior that gives a closed-form expression for the posterior is normally expected, particularly in high-dimensional parametric estimation and regression analysis. Otherwise, numerical integration methods such as Markov chain Monte Carlo *(*MCMC) needs to be applied. Conjugate prior means that the posterior distributions $p(\theta|data)$ are in the same family as the prior probability distribution $p(\theta)$. In our study we have used conjugate priors in our Bayesian inference.

It is useful to think of the construction of an empirical model as the process of combining historical and a-priori information. Alternative modelling techniques provide different a-priori information or different relative weights to sample the prior information. In-sample over fitting typically translates into poor forecasting performance. Bayesian methods address this problem; they make in-sample fitting less dramatic and improve out-of-sample performance.

We extend the models described in section 4.3 and apply multivariate versions of autoregressive and GARCH models, that is, vector autoregressive (VAR) and multivariate GARCH (MGARCH) models.

VAR models remove certain constraints arising from economics theory and are useful for multivariate analysis. But in large models with many parameters they have a problem with over-parameterization. Since the number of coefficients to be estimated quickly increases with the number of variables as well as the number of lags in the system, a moderate sized system can be highly over-parameterized relative to the number of observations. An over-parameterized unrestricted VAR model can explain data “too well”. It captures not only important features that are useful for forecasting, but also noisy features that merely reflect accidental or random relationships. Statistically, over-parameterization usually causes multicollinearity and loss of degrees of freedom, which lead to inefficient estimates and large out-of-sample forecasting errors. The solution can be found in two different methods – structural VAR and Bayesian VAR. A Bayesian VAR (BVAR) model, offers an intelligent way to overcome over-parameterization without relying on classical hypothesis testing. Due to the process of defining proper prior distributions for each parameter in the model, there will not be the situation where parameters are erroneously and coincidently considered as nonzero. Furthermore, BVARs enable forecasters to impose prior specifications through probabilistic terms in a fully transparent way. The means of prior distributions reflect forecasters’ prior beliefs and best guesses about the true values of unknown parameters. The variances reflect forecasters’ confidence on the prior means. Small prior variances indicate that forecasters believe the true values are not likely to deviate from their guesses (i.e., prior means), and vice versa. This standard specification procedure allows resulting forecasts to be reproduced. Moreover, BVARs generate complete multivariate density forecasts, by fully incorporating parameter uncertainty instead of simply using point estimates of parameters. Although the initial intention of BVAR models was to improve macroeconomic forecasts by Litterman (1979), they have evolved dramatically and are now used for a variety of purposes.

For the predictive modelling of volatility, the Bayesian inference approach is also adopted to the multivariate GARCH model. In our case, we have chosen to use the Dynamic Conditional Correlation (DCC) GARCH model built by Engle (2002). It considers correlation as a time-varying variable, which is the more realistic rendition of the multivariate GARCH models. Applying Bayesian inference to multivariate GARCH models is still a relatively new approach with only a few studies reporting results (Fioruci, Ehlers and Andrade Filho, 2014, Vrontos, Dellaportas and Politis, 2002).

The process by which we include news sentiment in the modelling framework is through the impact score derived in section 3.4. Instead of directly affecting the estimations of variables, the univariate prediction series estimated using the impact score is utilised to determine the distribution of the variable estimator. That is to say, the news data series is not required in the calculation of these multivariate models.

5.2 The Models

In this section we construct three models that enable us to obtain multivariate predictions for stock price return, volatility and liquidity. The choices of the respective models for these three characteristics of asset behaviour are explained in what follows. In the case of return, a multivariate linear model is applied. As for volatility, the multivariate case of the model that we use is the Multivariate GARCH (MGARCH) model (Bollerslev, Engle, Wooldridge, 1988), specifically the Dynamic Correlation Condition case (DCC) (Engle, 2002). As the estimation of such models are very costly due to the large number of parameters involved, we have taken the approach of using returns predicted by our news enhanced model as the input to calculating multivariate volatility. Lastly for liquidity, the adopted model takes a Vector Autoregressive (VAR) framework, which is common practice for multivariate analysis of asset characteristics (Reinsel, 2003; Lutkepohl, 1993).

Notation and Measures

We follow the notation set out in section 4.3 and provide a repeated description for convenience.

t – time period/time bucket.
i – lag in the time bucket.

Return, volatility, liquidity and news impact all use the same measures as explained in section 3.4 and 4.3.

Return

We choose a multivariate linear model consisting of two lags for asset return prediction, following the lag choice from section 4.3. Initial analysis of the stock returns data showed little correlation between the asset returns through the study of correlation plots. For this reason, a direct multivariate autoregressive model was not applied (otherwise known as vector autoregressive model).

Let $\boldsymbol{R_t}$ be a *d* x *1* vector of log-returns for *d* number of assets in period *t*,

$\boldsymbol{R_{t-i}}$ be *d* x *1* vectors of log-returns with lag *i* for *d* assets,

$\boldsymbol{A_i}$ be *d* x *d* coefficient matrices for the variables at lag period *i*,

$\boldsymbol{\eta_t}$ be the error term in time *t.*

Therefore, the returns expressed as $\boldsymbol{R_t}$ is given by the following multivariate linear equation,

$$\boldsymbol{R_t = A_0 + A_1 R_{t-1} + A_2 R_{t-2} + \eta_t} \qquad [5.1]$$

Volatility

Variance and standard deviation measure the spread of a distribution around its mean and is easily deduced from the covariance matrix of asset returns. The leading diagonal contains the variance values and square rooting those values give standard deviations. We have therefore adopted the DCC-GARCH model which is described below.

In general, denote $\boldsymbol{R_t}$ as the return series and assume that (Bauwens et al., 2006):

$$\boldsymbol{R_t = \mu_t(\theta) + \varepsilon_t} \qquad [5.2]$$

and the $\boldsymbol{\varepsilon_t}$ term equates to

$$\boldsymbol{\varepsilon_t = H_t^{1/2}(\theta) z_t} \qquad [5.3]$$

where $\boldsymbol{\theta}$ is the vector of parameters,

$\boldsymbol{\mu_t}(\boldsymbol{\theta})$ is a *d* x *1* vector of conditional means,

$\boldsymbol{H_t}(\boldsymbol{\theta})$ is a *d* x *d* matrix of conditional variances associated with the log return $\boldsymbol{R_t,}$

$\boldsymbol{z_t} \sim \boldsymbol{N}(0, \boldsymbol{I})$.

Then $\boldsymbol{H_t}$ (omitting $\boldsymbol{\theta}$ thereafter for convenience), the conditional covariance matrix, is expressed as

$$\boldsymbol{H_t} = \boldsymbol{D_t P_t D_t} \qquad [5.4]$$

where $\boldsymbol{D_t} = diag(h_{11t}^{1/2} \dots h_{ddt}^{1/2})$,

h_{iit} can be any univariate GARCH model for d number of assets,

$\boldsymbol{P_t} = diag(q_{11t}^{1/2} \dots q_{ddt}^{1/2})\boldsymbol{Q_t} diag(q_{11t}^{1/2} \dots q_{ddt}^{1/2})$,

$\boldsymbol{Q_t} = (q_{ijt})$ is the *d* x *d* symmetric positive definite matrix defined as

$$Q_t = (1 - \alpha - \beta)\bar{Q} + \alpha u_{t-1}u'_{t-1} + \beta Q_{t-1}$$

with $u_{it} = r_{it}/\sqrt{h_{iit}}$.

$\boldsymbol{P_t}$ represents the dynamic (time-varying) co-movements between assets. α depicts the impact of past standardised shocks and β measures the impact of lagged dynamic conditional correlations on those dynamic conditional correlations at time *t*. Both of these parameters are non-negative and a necessary condition that $\alpha + \beta < 1$ must hold. $\overline{Q} = n^{-1}\sum_{t=1}^{n} u_t u'_t$ is the unconditional correlation matrix of $\boldsymbol{R_t}$**.**

This multivariate GARCH model allows one-step-ahead volatility values to be forecasted. This is implemented using an iterative process. With such inputs we are able to calculate $\boldsymbol{Q_t}$ **.** Bayesian inference is adopted here to incorporate prior information and the likelihood function to inference $\boldsymbol{P_t}$ (and therefore $\boldsymbol{H_t}$).

Given the conjugate prior outlined in section 5.3, we are also able to derive $\boldsymbol{P_t}$ as a Wishart distribution as follows (Brown, Vannucci and Fearn, 1998):

$$\boldsymbol{P_t} \sim W(d + n, \boldsymbol{D_t V D}_t^{-1}) \qquad [5.5]$$

where $\boldsymbol{V} = \ \mathbf{W^{-1}} + \mathbf{S} + \frac{n}{1+n}(\boldsymbol{m} - \boldsymbol{\bar{x}})(\boldsymbol{m} - \boldsymbol{\bar{x}})^{\mathrm{T}}$,

d is the number of assets,

n is the length of sample data used,

$\boldsymbol{m}$ is the mean of returns estimated from a news enhanced model,

$\boldsymbol{\bar{x}}$ is the mean of returns estimated from a market data only model,

$\mathbf{W^{-1}}$ is the covariance matrix of volatility estimated with news sentiment data,

$\mathbf{S} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^{\boldsymbol{T}}$, x_i are data points from return series estimated using only market data,

$\boldsymbol{D_t}$ as above.

It is possible to derive this because $\boldsymbol{H_t}$is also a Wishart distribution. Therefore, with the distribution of $\boldsymbol{P_t}$ determined, we are able to update or estimate $\boldsymbol{P_t}$ by the expectation of the posterior distribution, which is given by:

$$E\lfloor \boldsymbol{P_t} \rfloor = (d+n) * \lfloor \boldsymbol{D_t V D_t^{-1}} \rfloor \qquad [5.6]$$

For further details see section 5.3. It should be noted that DCC-GARCH models are relatively insensitive to the choice of the univariate model specification, so whether a straight forward GARCH model is used or a variation will not alter the multivariate results extensively (Cappiello, Engle and Sheppard, 2006).

Liquidity

A VAR model of lag three is used for the prediction of liquidity (using the bid-ask spread as a proxy). Taking a direct extension of the univariate liquidity prediction model, we include the correlation between liquidity values of different assets, contrasting the case of the asset return model.

Let, $\boldsymbol{S_t}$ be a *d* x *1* vector of bid-ask spreads for *d* number of assets,
$\boldsymbol{B_i}$ be *d* x *d* coefficient matrices with *i* = 1,2,3,
$\boldsymbol{e}_t$ be the error term

Therefore, the liquidity (bid-ask spread) expressed as $\boldsymbol{S_t}$ is given by

$$\boldsymbol{S}_t = \sum_{i=1}^{3} \boldsymbol{B}_i \boldsymbol{S}_{t-i} + \boldsymbol{e}_t \qquad [5.7]$$

5.3 Prior Selection and Posterior Distributions

We follow the arguments and concepts of the Black-Litterman model (1991) and regard news sentiment as extra knowledge to be considered in the model estimation (it is only considered externally). That is to say, neither the impact score nor the news sentiment data is used in the multivariate predictive models or in Bayesian inference. Instead, the univariate predicted series of return, volatility and liquidity obtained in Chapter 4 are used in the calculation of prior and posterior distributions. It is in this way that news sentiment is considered as additional information and incorporated in the modelling process. Similar to the experiments carried out in the univariate case, we also set out to compare the performance of the multivariate predictive models considering news sentiment against those that only consider market data. This is achieved through comparison of uninformative priors and informative priors, where the uninformative priors only use market data.

Return Model

For the asset return model, we assume a Gaussian distribution for $\boldsymbol{R_t}$.

$$\boldsymbol{R_t} \mid \mu, \Phi \sim \mathcal{N}(\mu, \Phi^{-1})$$

where Φ is d x d matrix. We aim to infer μ and Φ by Bayesian inference based on the likelihood and prior, where the likelihood comes from market data and the prior comes from the news sentiment data.

First we consider the simple case where the d assets are assumed to be independent, which means that $\Phi^{-1} = \sigma^2 \mathbf{I}$ and σ^2 is known. Then we have a conjugate prior for μ as normal and is written as:

$$\mu | S_0 \sim N(\mu_0, S_0)$$

Due to the selection of conjugate priors, closed form equations are provided to calculate the posterior distribution. The posterior of μ is given by:

$$P(\mu \,| y) \sim \mathcal{N}(\mu, S)$$

where y denotes data and

$$S^{-1} = S_0^{-1} + \frac{1}{\sigma^2} X^T X \qquad [5.8]$$

$$\mu = S(S_0^{-1}\mu_0 + \frac{1}{\sigma^2} X^T y) \qquad [5.9]$$

X is a matrix based on lag returns. Clearly μ is also the posterior mean in this case. Furthermore, the posterior mean combines both market data and news sentiment data. A flat (uninformative) prior is also constructed. In contrast

$$P(\mu) \propto 1.$$

Then the posterior distribution is normal and derived as:

$$P(\mu\,|y) = \mathcal{N}((X^T X)^{-1} X' y, \sigma^2 \mathbf{I}) \qquad [5.10]$$

which is the same as the frequency result.

Volatility Model

Starting from the assumption of our multivariate GARCH model (see section 5.2) we aim to infer $\boldsymbol{H_t}$**.** Under the Bayesian structure, a Normal-Wishart prior is applied, which is a conjugate prior. Taken from Bayesian statistics, the conjugate prior for the mean vector of a multivariate normal distribution is another multivariate normal distribution, and the conjugate prior for the covariance matrix is an inverse-Wishart distribution. Let $\boldsymbol{\Phi} = \boldsymbol{H_t^{-1}}$, we obtain a Wishart distribution for $\boldsymbol{\Phi}$. Therefore,

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Phi^{-1}})$$
$$\boldsymbol{\Phi} \sim \mathcal{W}\ (d, \boldsymbol{W})$$

where d is the number of assets,

$\mathbf{W^{-1}}$ is the covariance matrix of volatility estimated with news sentiment data,

$\boldsymbol{m}$ is the mean of returns estimated from a news enhanced model.

Thus we are able to use the analytical (closed form) formula to estimate the posterior mean, which is the estimator. According to Bayes' theorem the posterior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ are given by

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \boldsymbol{y} \sim \mathcal{N}\left(\frac{\boldsymbol{m} + \boldsymbol{\bar{x}}}{1+n}, \frac{1}{1+n} \boldsymbol{\Phi}^{-1}\right)$$

$$\boldsymbol{\Phi} \mid \boldsymbol{y} \sim \mathcal{W}\left(d + n, \mathbf{W^{-1}} + \mathbf{S} + \frac{n}{1+n}(\boldsymbol{m} - \boldsymbol{\bar{x}})(\boldsymbol{m} - \boldsymbol{\bar{x}})^{\mathrm{T}}\right)$$

where n is the sample size used to determine the prior,

$\mathbf{S} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^{\boldsymbol{T}}$, x_i are data points from return series estimated using only market data,

$\boldsymbol{\bar{x}}$ is the mean of returns estimated from a market data only model.

For convenience, we denote the second distribution parameter of the posterior Wishart distribution as $\mathbf{V}$ (this is used to derive equation 5.5). Evidently the Wishart distribution for the posterior is different from the Wishart distribution for the prior. Here the incorporation of news sentiment data can be identified. In the calculation of the covariance matrix in the posterior distribution both of the return series are used (the series estimated using market data only and the series from market data plus the impact score). Finally, taking the posterior mean to be the estimator for $\boldsymbol{H_t}$ we have

$$E[\boldsymbol{H_t}] = (d+n) * \boldsymbol{V} \qquad [5.11]$$

where $= \ \mathbf{W^{-1}} + \mathbf{S} + \frac{n}{1+n}(\boldsymbol{m}-\overline{\boldsymbol{x}})(\boldsymbol{m}-\overline{\boldsymbol{x}})^{\mathrm{T}}$ **.**

The limiting prior for the conjugate multivariate normal is taken to be our non-informative prior. It has the form $p(\boldsymbol{\Phi}) \ \propto |\boldsymbol{\Phi}|^{\frac{k+1}{2}}$**.** Hence the posterior distribution is

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \boldsymbol{y} \sim \mathcal{N}\left(\overline{\boldsymbol{x}}\,, \frac{1}{n}\boldsymbol{\Phi}^{-1}\right)$$
$$\boldsymbol{\Phi} \mid \boldsymbol{y} \sim \mathcal{W}(n-1, \boldsymbol{S})$$

This does not involve the news sentiment series at all.

Liquidity Model

For convenience, we form a matrix $\mathcal{S} = (\mathbf{S_1}, \mathbf{S_2}, \dots, \mathbf{S_n})$. Assume $\mathcal{S}$ follows a multivariate normal.

$$\mathcal{S}|\boldsymbol{\Gamma}, \boldsymbol{\Psi} \sim \mathcal{N}(\boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Omega})$$

or written as $\boldsymbol{\Gamma}|\ \boldsymbol{\Psi} \sim \mathcal{N}(\boldsymbol{\Gamma}, \boldsymbol{\Psi}\otimes\boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = (\boldsymbol{Z^T Z})^{-\mathbf{1}}$ with $\mathbf{Z}$ being a matrix of lagged liquidity terms, and $\boldsymbol{\Psi}$ is the covariance matrix for each $\mathbf{S_t}$**.** Again we use the conjugate prior for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$.

$$\boldsymbol{\Gamma} \mid \boldsymbol{\Psi} \sim \mathcal{N}(\underline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}})$$
$$\boldsymbol{\Psi} \sim i\mathcal{W}(\underline{\mathbf{S}}, \underline{\upsilon}\ )$$

where $\underline{\boldsymbol{\Gamma}}$**,** $\underline{\mathbf{S}}$, $\underline{\upsilon}$ and $\underline{\boldsymbol{\Omega}}$ are calculated from the liquidity series that uses news sentiment data. The posterior distributions are derived as below.

$$\boldsymbol{\Gamma} \mid \boldsymbol{\Psi}\,, \mathbf{y} \sim \mathcal{N}(\overline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \overline{\boldsymbol{\Omega}})$$
$$\boldsymbol{\Psi}|\ \mathbf{y} \sim i\mathcal{W}(\underline{\mathbf{S}}, \underline{\upsilon})$$

$$\overline{\boldsymbol{\Omega}}^{-1} = \underline{\boldsymbol{\Omega}}^{-1} + \mathbf{Z'Z} \qquad [5.12]$$
$$\overline{\boldsymbol{\Gamma}} = \overline{\boldsymbol{\Omega}}(\underline{\boldsymbol{\Omega}}^{-1}\underline{\boldsymbol{\Gamma}} + \mathbf{Z'Z}\hat{\boldsymbol{\Gamma}}) \qquad [5.13]$$

with $\hat{\mathbf{\Gamma}} = (\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{Y}$ and $\mathbf{S} = (\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}})'(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{\Gamma}})$. Accordingly the normal-diffuse prior is used for considering market data only. It takes the form:

$$p(\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-(m+1)/2}$$

Using this prior in the Bayesian VAR model, we obtain the posterior distribution as the following

$$\mathbf{\Gamma}|\, \boldsymbol{y}, \mathbf{\Psi} \sim \mathcal{N}(\bar{\gamma}, \bar{\Sigma}_{\gamma})$$
$$\mathbf{\Psi}|\, \boldsymbol{y}, \mathbf{\Gamma} \sim i\mathcal{W}(\bar{\boldsymbol{S}}, \bar{v})$$

$$\bar{\mathbf{\Sigma}}_{\gamma} = (\underline{\mathbf{\Sigma}}_{\gamma}^{-1} + \mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z})^{-1} \quad [5.14]$$
$$\bar{\gamma} = \bar{\mathbf{\Sigma}}_{\gamma}\left[\underline{\mathbf{\Sigma}}_{\gamma}^{-1}\underline{\gamma} + (\mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z})\hat{\gamma}\right] \quad [5.15]$$

where $\hat{\gamma} = [\mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}]^{-1}(\mathbf{\Psi}^{-1} \otimes \mathbf{Z}')\boldsymbol{y}$.

Only the distribution of γ is of importance in our case as the estimator we want is given by the mean of this distribution.

5.4 Data

In general, if a trader wishes to trade in a financial stock, for example, then they would execute a transaction in one or a collection of stocks within the same sector. Therefore, with this in mind, we have chosen five stocks from the Finance industry: AIG, Bank of America, Barclays, HSBC and JP Morgan. The fact that all these stocks belong to the same sector ensures that a correlation exists within the portfolio. As in the univariate case, market and news sentiment data are extracted for the year 2008 to fit the models, and then the proceeding three months of January-March 2009 are used for out-of-sample testing. Throughout this work, the impact score described in section 3.4 is used as the data series for news sentiment.

5.5 Computational Results and Validation

To compare the performance of these news enhanced Bayesian models, direct comparison is made between the models described in Section 5.2 and the equivalent predictive models that only use market data i.e. adopts a flat prior. This flat prior provides no additional information or knowledge but follows the same estimation procedure as those described in section 5.3. We use market data only models as a benchmark to compare the performance against our multivariate models. From this, the specific impact of news sentiment on prediction power is revealed. Using in sample data of 2008 for five financial assets, we estimate the coefficient matrices for the stock price return model that forms its prior based on the news sentiment data series (see equation 5.1). They are:

$$\boldsymbol{A_0} = \begin{bmatrix} -1.96e^{-5} \\ -6.36e^{-6} \\ -3.73e^{-6} \\ -8.46e^{-7} \\ -2.20e^{-6} \end{bmatrix}, \boldsymbol{A_1} = \begin{bmatrix} 1.72e^{-1} \\ 2.93e^{-3} \\ -2.28e^{-3} \\ 3.48e^{-3} \\ -7.51e^{-3} \end{bmatrix}, \boldsymbol{A_2} = \begin{bmatrix} 1.04e^{-2} \\ 1.23e^{-2} \\ -2.92e^{-4} \\ 3.42e^{-3} \\ -1.62e^{-2} \end{bmatrix}$$

The coefficient matrices estimated from market data only (flat prior) are:

$$\boldsymbol{A_0} = \begin{bmatrix} -8.96e^{-6} \\ -2.80e^{-6} \\ -2.50e^{-6} \\ -8.36e^{-7} \\ -8.70e^{-7} \end{bmatrix}, \boldsymbol{A_1} = \begin{bmatrix} 4.21e^{-2} \\ -1.85e^{-2} \\ -4.25e^{-1} \\ -6.06e^{-1} \\ 1.60e^{-1} \end{bmatrix}, \boldsymbol{A_2} = \begin{bmatrix} -2.85e^{-3} \\ 6.31e^{-3} \\ -1.63e^{-1} \\ -2.86e^{-1} \\ -2.71e^{-2} \end{bmatrix}$$

Evidently, these two sets of matrices are not equal and therefore will have different results for the prediction of returns.

As for the multivariate volatility model, we present the estimated values for the measures of impact from past standardised shocks (α) and from lagged dynamic conditional correlations (β) – see table 5.1. This was calculated using the package 'CCgarch' in the software R. The necessary condition of $\alpha + \beta < 1$ holds for both models.

DCC-GARCH Model	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
Market data only	$3.40640e^{-3}$	0.86363
Market and News data	$5.99723e^{-4}$	0.99733

Table 5.1: Estimated values of α and β parameters for the DCC-GARCH model in two cases – with only market data and including news sentiment data with market data.

The coefficient matrix to be estimated for the prediction model of liquidity (by using the bid-ask spread as a proxy) with the consideration of news sentiment is $\mathbf{B}_{\text{News}} = (\mathbf{B_1}, \mathbf{B_2}, \mathbf{B_3})'$ (see equation 5.7). The sample calculations resulted this to be

$$\mathbf{B}_{\text{News}} = \begin{bmatrix} 1.00216 & -0.13193 & -0.02597 & 0.29831 & -0.09181 \\ 1.69672 & 2.10184 & 0.18117 & -2.14771 & 0.62422 \\ -1.65849 & -1.07422 & -0.19660 & 0.86639 & -0.82184 \\ -0.00730 & -0.00531 & -0.00096 & 0.31607 & -0.00326 \\ 0.46910 & 0.35984 & 0.06573 & -0.77433 & 1.36271 \\ 0.37419 & 0.02608 & 0.00970 & -0.10406 & 0.03520 \\ -1.72855 & -0.36563 & -0.15338 & 1.83415 & -0.52787 \\ -0.72300 & -0.80197 & 0.43951 & 0.72228 & -0.57991 \\ -0.00240 & -0.00174 & -0.00033 & 0.52751 & -0.00125 \\ -0.53551 & -0.41238 & -0.07532 & 0.87568 & 0.20733 \\ -0.39206 & -0.06185 & -0.01368 & 0.16110 & -0.04981 \\ 1.65616 & 0.69656 & 0.19438 & -2.31878 & 0.67416 \\ -0.99812 & -0.81910 & -0.51335 & 1.61581 & -0.79118 \\ -0.00715 & -0.00497 & -0.00089 & 0.23906 & -0.00299 \\ 0.62669 & 0.48682 & 0.08965 & -1.05060 & -0.06043 \end{bmatrix}$$

Similarly, the coefficient matrix for the BVAR model using market data only gives the following estimator

$$\mathbf{B}_{\text{M}} = \begin{bmatrix} 0.04544 & 0.00645 & 9.94e^{-6} & 0.00012 & 0.00057 \\ 0.01829 & 0.13064 & 2.46e^{-5} & -0.00077 & 0.00040 \\ -0.59828 & -0.63509 & 0.21269 & -0.01397 & -0.64502 \\ 1.17e^{-5} & -7.66e^{-5} & 1.91e^{-8} & 0.00177 & 8.61e^{-6} \\ 0.00460 & 0.001738 & -3.81e^{-5} & 0.00070 & 0.09978 \\ 1.78719 & -0.01761 & -1.51e^{-5} & 4.74e^{-5} & -0.00077 \\ -0.04747 & 1.87393 & -2.56e^{-5} & 0.00226 & -0.00147 \\ -0.18676 & -0.31596 & 0.36113 & -0.00706 & -0.36924 \\ 3.01e^{-6} & 9.83e^{-5} & -8.49e^{-7} & 1.15984 & -2.11e^{-5} \\ -0.00416 & -0.00260 & 0.00011 & -0.00164 & 1.91329 \\ -0.77110 & 0.01199 & 5.99e^{-6} & -6.15e^{-5} & 0.00027 \\ 0.03406 & -0.90764 & -2.54e^{-6} & -0.00151 & 0.00125 \\ -0.69388 & -0.70095 & -0.57381 & 0.04259 & -0.77381 \\ 0.00017 & 9.36e^{-6} & 3.49e^{-6} & 0.33082 & 2.17e^{-5} \\ 0.00202 & 0.00189 & -7.51e^{-5} & 0.00092 & -0.90284 \end{bmatrix}$$

Once again these matrices are numerically different indicating that the addition of news sentiment affects the prediction of liquidity, and hence will give differing predictive performance to just market data. Next we evaluate this difference in performance.

To understand the extent to which news sentiment impacts these models, the fitted models obtained in sample are used to predict out-of-sample values. This period has been chosen to be a duration of 3 months straight after the in sample period, i.e. January-March 2009. Predictions are made for one-step ahead, so that is the next minute of trading. The performance of the models is measured according to the accuracy of the prediction against the true values. Below we present several statistical measures describing the estimation accuracy by the two sets of models.

Firstly, for stock price return, boxplots for the difference in residuals between the two types of models are provided for each of the five assets (see figure 5.1). Through observation of this graph it can be concluded that there is little difference in the prediction power of the multivariate model that incorporates news sentiment and one that does not. In other words, no added value is brought about by including news sentiment in stock price return prediction. This is demonstrated in the dispersion of the difference in residuals with the median for all assets grouped around zero, except for one – JP Morgan. Figure 5.2 displays a magnified plot excluding JP Morgan to show the exact quartiles of the data series. Observing with the naked eye it can be identified that Bank of America is the only asset that has a median value which is positive. The financial stock of JP Morgan exhibits a distinct behaviour in contrast to the others in that its median of difference in residuals is highly negative, indicating a worse performance by the news enhanced model than the market data only model. Moreover, JP Morgan also has the greatest range in data values with the whiskers of the plot representing the highest and lowest data points falling within 1.5*IQR.

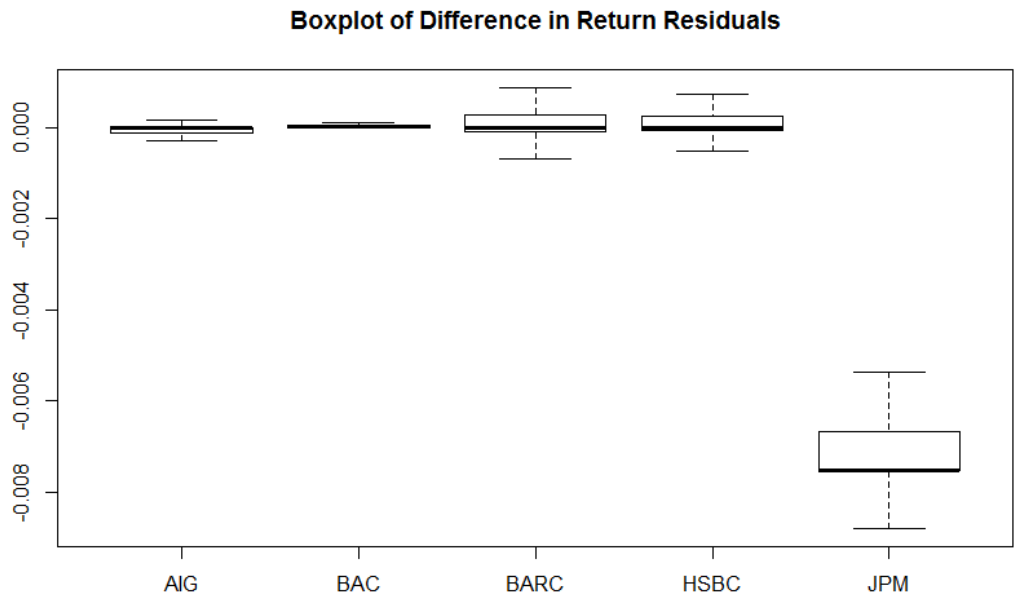


Figure 5.1: Boxplot of difference in residuals between a multivariate linear model for stock price return that incorporates news sentiment and one that does not, for five assets (AIG, Bank of America, Barclays, HSBC and JP Morgan). The difference is taken as the market data only model minus the news enhanced model. Outliers are not included in the plot for clarity reasons.

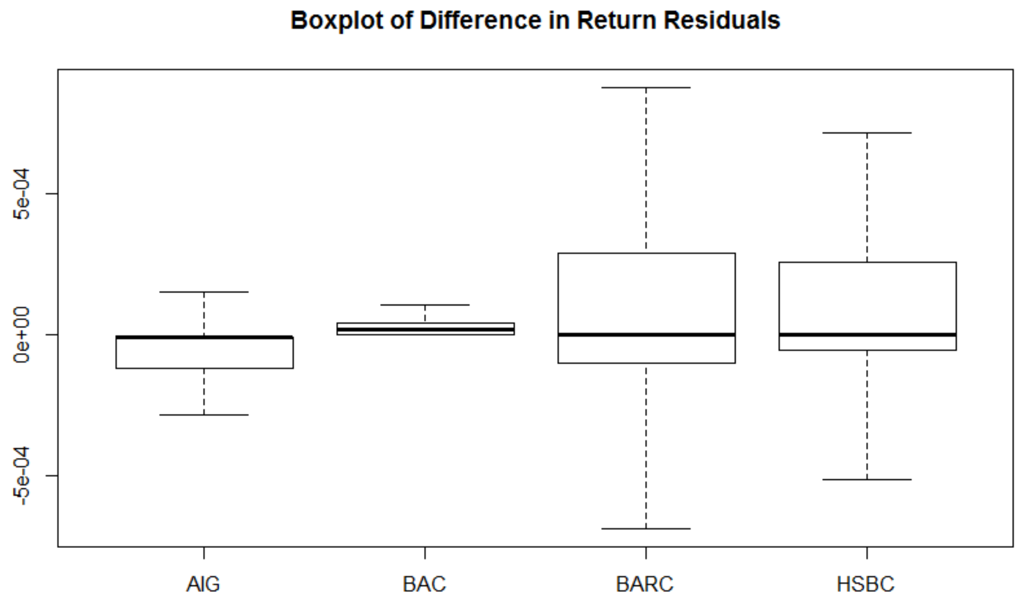


Figure 5.2: Boxplot of difference in residuals between a multivariate linear model for stock price return that incorporates news sentiment and one that does not, for four assets (AIG, Bank of America, Barclays and HSBC). The difference is taken as the market data only model minus the news enhanced model. Outliers are not included in the plot for clarity reasons.

Secondly, the boxplot for difference in residuals for the volatility models are given in figure 5.3. As with the case for return, the volatility prediction performance of one asset is particularly noticeable from the rest - AIG. From the boxplot it can be seen that it possesses the most negative series and also the widest range in data points represented by the long whiskers extending from the box, in comparison to the other assets. Such negativity indicates that generally, the predictions by the market data only multivariate GARCH model are closer to the true value than those predicted by the model that incorporates news sentiment. The other four remaining assets all exhibit comparable features, that is their median value of difference in residuals are all close to zero and their lowest data point that falls within 1.5*IQR is smaller than -$7.00e^{-6}$ (see figure 5.4). Furthermore, there is an obvious negative skew to these differenced residuals with the third quartile value at a far greater distance from the median compared with the first quartile. The longer extension of the whisker in the negative direction also indicates a negative skew to the data.

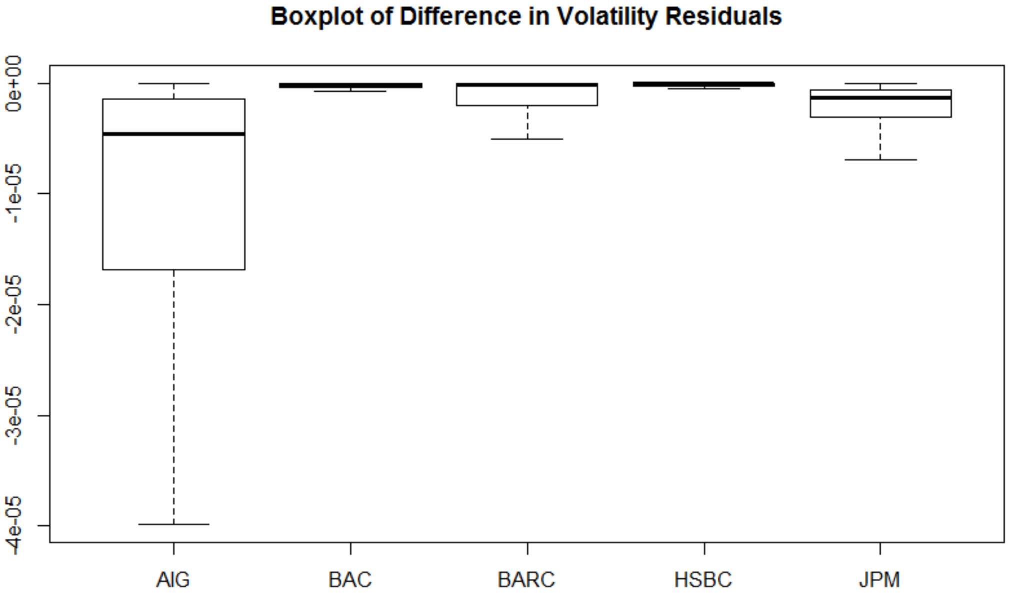


Figure 5.3: Boxplot of difference in residuals between a DCC-GARCH model for volatility that incorporates news sentiment and one that does not, for five assets (AIG, Bank of America, Barclays, HSBC and JP Morgan). The difference is taken as the market data only model minus the news enhanced model. Outliers are not included in the plot for clarity reasons.

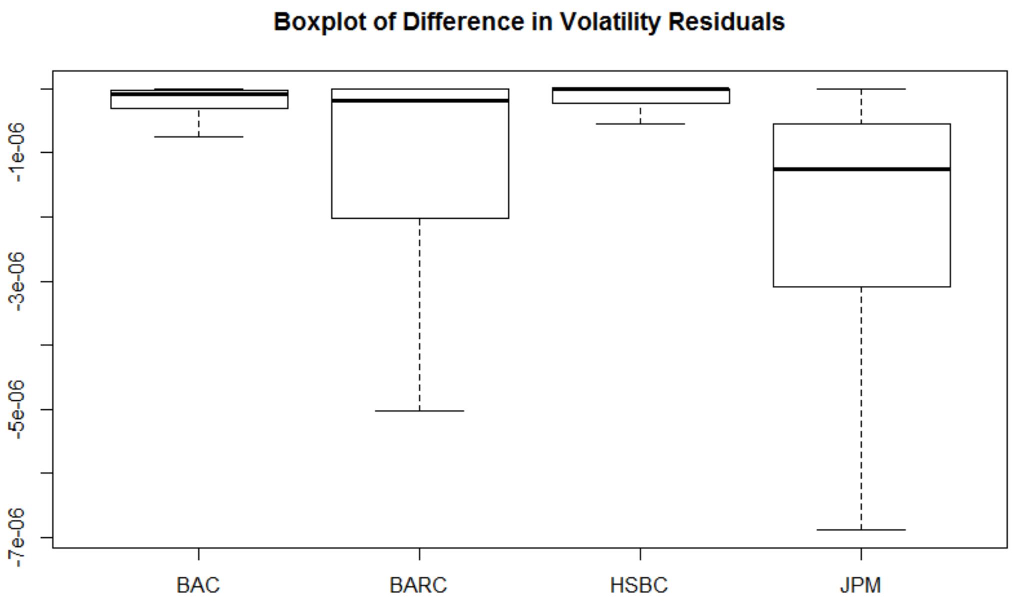


Figure 5.4: Boxplot of difference in residuals between a DCC-GARCH model for volatility that incorporates news sentiment and one that does not, for four assets (Bank of America, Barclays, HSBC and JP Morgan). The difference is taken as the market data only model minus the news enhanced model. Outliers are not included in the plot for clarity reasons.

Lastly, the results for liquidity prediction vary significantly as suggested by the boxplot in figure 5.5. Out of all five selected assets, both positive and negative median values appear meaning that predictions of liquidity values (by using the bid-ask spread as a proxy) are improved with the consideration of news sentiment for some assets but not others. Specifically, AIG and Barclays show positive results but Bank of America and JP Morgan show negative results. Similar to the results for volatility, HSBC shows little dispersion in the difference of residuals for liquidity modelling and the same observation applies to AIG for predictions of return and liquidity. However, compared to the previous two sets of predictive models, these liquidity predictions produce large values of difference in residuals. Here, the range is between -0.1 and 0.1 whereas the former results are all within a range of -0.001 and 0.001 (bar the exception of JP Morgan return predictions).

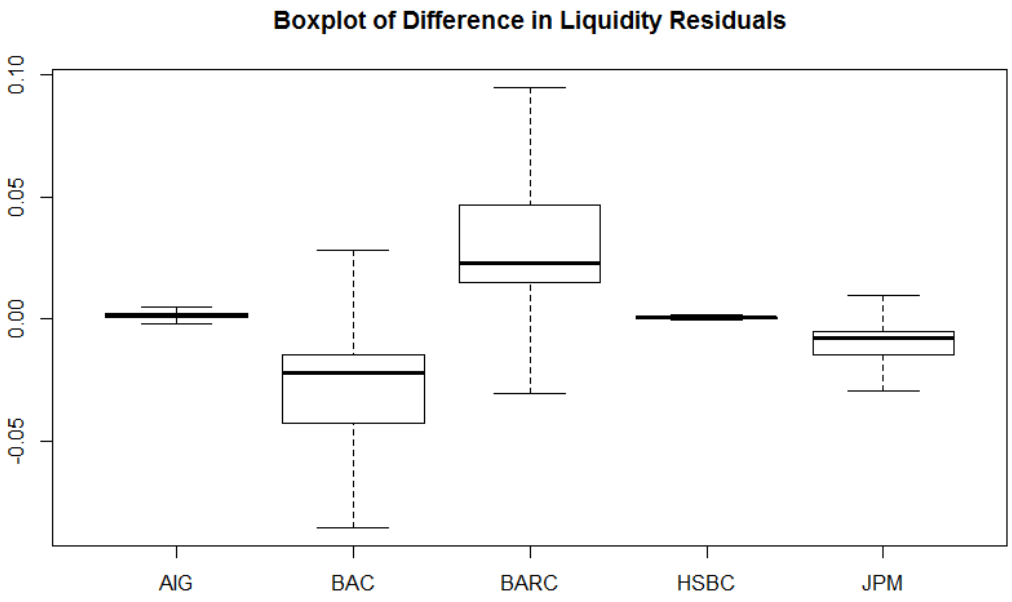


Figure 5.5: Boxplot of difference in residuals between a VAR model for liquidity that incorporates news sentiment and one that does not, for five assets (AIG, Bank of America, Barclays, HSBC and JP Morgan). The difference is taken as the market data only model minus the news enhanced model. Outliers are not included in the plot for clarity reasons.

In order to be able to deduce some conclusions with more certainty, we calculated additional statistical measures, namely root mean squared errors (RMSE) and mean absolute errors (MAE) for the residual time series of each asset. The difference is taken between the errors for the market only model and the news enhanced model to clearly show which performs best. A positive value represents a better prediction by the model that considers news sentiment and a negative value suggests otherwise. Table 5.2 displays the RMSE values, which depicts an improved prediction in volatility by the news enhanced model with the majority of differences giving positive results, albeit values that are very close to zero. However, liquidity results are not so promising with all differences presenting negative values except one – HSBC. This is consistent with the interpretation from the boxplot of liquidity residuals. Return prediction performance is inconclusive as some asset returns are better predicted by the market data only model and some are better predicted by the multivariate linear model that incorporates news sentiment. Exactly the same conclusions are made by looking at the MAE values (see table 5.3).

RMSE		**AIG**	**BAC**	**BARC**	**HSBC**	**JPM**
Return	*Without sentiment*	0.00370	0.00240	0.00508	0.01587	0.00127
	With sentiment	0.00387	0.00233	0.00498	0.01785	0.00772
	Difference	-0.0017	0.00007	0.00010	-0.00198	-0.00645
Volatility	*Without sentiment*	0.00271	0.00229	0.00418	16.10897	0.00276
	With sentiment	0.00260	0.00229	0.00414	16.65892	0.00276
	Difference	0.00011	0.00000	0.00004	-0.54995	0.00000
Liquidity	*Without sentiment*	0.00831	0.00944	0.05588	0.03350	0.00097
	With sentiment	0.01117	0.08985	0.14200	0.02300	0.03282
	Difference	-0.00286	-0.08041	-0.08412	0.01050	-0.03185

Table 5.2: Root mean squared errors (RMSE) for the prediction of stock price return, volatility and liquidity for each asset (AIG, Bank of America, Barclays, HSBC and JP Morgan). A separation is made between the multivariate model that incorporates news sentiment and the one without. The difference value indicates which model performs better. Numbers are given to 5 decimal points.

Mean Absolute Error		**AIG**	**BAC**	**BARC**	**HSBC**	**JPM**
Return	*Without sentiment*	0.00062	0.00108	0.00135	0.00096	0.00061
	With sentiment	0.00754	0.00105	0.00118	0.000762	0.00754
	Difference	-0.00692	0.00003	0.00017	0.000199	-0.00692
Volatility	*Without sentiment*	0.00250	0.00209	0.00332	0.95309	0.00262
	With sentiment	0.00249	0.00209	0.00330	1.09715	0.00261
	Difference	0.00001	0.00000	0.00002	-0.14406	0.00001
Liquidity	*Without sentiment*	0.00209	0.00138	0.03715	0.00261	0.00059
	With sentiment	0.00211	0.03888	0.07902	0.00176	0.01356
	Difference	-0.00002	-0.0375	-0.04187	0.00085	-0.01297

Table 5.3: Mean absolute error (MAE) for the prediction of stock price return, volatility and liquidity for each asset (AIG, Bank of America, Barclays, HSBC and JP Morgan). A separation is made between the multivariate model that incorporates news sentiment and the one without. The difference value indicates which model performs better. Numbers are given to 5 decimal points.

5.6 Summary

In this chapter we have considered in a multivariate setting the prediction of return, volatility and liquidity for a small universe of trading assets. This requires multivariate techniques that include the consideration of correlation between assets. This justifies our selection of five assets all belonging to the Finance industry so that correlations are guaranteed to be strong. We also note that this is a natural way traders rebalance (holds and liquidates positions) a trading portfolio. Bayesian inference is adopted because it combines information from the likelihood and the prior. We first propose the inclusion of news sentiment data for priors. This is the proper way to use news information in this context because it occurs at the same time as market information. This is a better method than historical data based priors because historical data may be out of date or inconsistent with current data. By considering news sentiment metadata in the formation of priors, we acknowledge the additional information supplied and use this to construct Bayesian prediction models. The results of our empirical tests show that some improvements can be made in the prediction of volatility once news sentiment is considered, which is consistent with what was found in the univariate case. Another result that is similar to single asset predictions is stock price return, specifically the inability to better the predictions made by the market data only model. However, when modelling liquidity in a multivariate framework it seems that the added value contained in the news enhanced model seen in the univariate case is lost. This, however, does not imply that the incorporation of news sentiment data or Bayesian inference is not useful. Our contention is that the results point us to consider an alternative Bayesian inference model (other than the normal distribution); but this is outside the scope of this thesis and can be explored in future work. An additional feature that may bring more insight into this work is the consideration of prediction performance of a whole portfolio. Instead of reporting the asset behaviour of individual companies, the return, volatility and liquidity variables of a portfolio built from all 5 assets may better indicate the effect of news sentiment on asset behaviour prediction.

Chapter 6

Conclusions

6.1 Summary

In this thesis we have explored how news stories as events can be incorporated in predictive models in the domain of finance. We have studied and reported on research literature in this area which describes the evolution of proxies for sentiment and how they impact asset behaviour. We have considered briefly how news stories are converted to quantified measures. The power of unstructured text and sentiment quantification is an important research topic in the field of machine learning, however, the focus of our research has been on the application of sentiment measures as supplied by trusted content suppliers (see Appendix A and Appendix B).

Since the research results reported in this thesis are of value to traders and risk officers, we have reviewed the field of market microstructure and automated trading. In our predictive data models we have used minute bar data and constructed predictive models whose results find use in automated high frequency (intraday) trading. We have further considered the topic of liquidity in some depth since we have introduced liquidity as an additional parameter in the description of asset behaviour.

To set the context of the predictive models reported in this thesis – the input data, the scope and purpose (the focus) – we consider in Chapter 3 news metadata and some attributes such as entity recognition, relevance, novelty and sentiment score. Earlier researchers have only used sentiment scores in analysing the impact of news on financial instruments. We have, however, formally defined an impact measure which takes into account (i) news flow and (ii) the decaying effect of news sentiment.

News data arrives asynchronously during the day and market data is available at various frequencies (tick data, minute bar, end of day setting). Our predictive models use these two time series data, namely, asynchronous daily news metadata and minute bar market data. The predictions made by our news enhanced GARCH (1,1) and AR(3) models for volatility and liquidity, respectively, showed superior performance.

The volatility results displayed substantial improvements to a market data only model, which is consistent with previous research findings. The improved predictions of liquidity do not compare to those seen in volatility but does show better performance than stock price return predictions, which are not improved at all.

In the multivariate models which are used to predict the behaviour of a collection of assets belonging to the same industry, correlation effects deteriorate the superior - prediction performance seen in the univariate setting. The most promising results are seen from the DCC-GARCH model that predicts multivariate volatility but performance is no longer highly significant. When modelling liquidity using a Bayesian VAR model no further value can be added by incorporating the impact score. We observe that the multivariate predictive model for return is unable to produce improved predictions.

6.2 Conclusions and Contributions

The study of our thesis is motivated by two research problems which are determining trading strategies and quantifying trading risk. In addition to return and volatility, we have considered liquidity such that taken together (return, volatility and liquidity) these provide a complete characterisation of asset behaviour. Our use of a predictive model for liquidity provides a framework for capturing the liquidity risk for executable trades. During our empirical investigation to discover a relationship between news arrival and asset price, we gained some insight into how news sentiment and asset prices are related. In section 3.2 and 3.4 we report the results of two χ^2 tests which examine the connections between news sentiment and asset price. We observe that there exists a stronger relationship between impact score and prices compared to sentiment of individual news events and price. In the preliminary stages of our investigation, we set out to predict asset behaviour by using these news sentiment scores calculated and supplied by data vendors. However, the performance of models which used individual news sentiment scores was weak leading us to abandon a direct application of these sentiment scores and explore alternative measures. Thus the negative results of this investigation led us to find a positive way of quantifying the relationship; hence we derived the impact score. The impact score is in some sense a natural and intuitive measure as this takes into account news flow and its effect on market movements. Our two sets of predictive models use this new measure, namely, the impact score which takes into account (i) the volume of news and (ii) the decaying effect of news sentiment. In such a manner we derive the impact of aggregated news events for a given asset. The derivation of the impact measure and the characterisation of asset behaviour by introducing liquidity are two innovations reported in this thesis and are contributions to knowledge.

A critical evaluation of the results of our univariate and multivariate models lead us to conclude that univariate predictions are far better than multivariate predictions. However, multivariate models play an important role by providing a blueprint for scenario generation which can be used for constructing trading strategies and quantifying trading risk. The Black-Litterman model (1991) is widely studied and reported because of its inclusion of human judgement (expert knowledge) in the prediction of asset behaviour. We have taken the underlying principle of Black-Litterman (1991) and constructed Bayesian multivariate predictive models such that domain expertise is incorporated to improve predictions. This is another novel aspect of our work and a further contribution to knowledge.

6.3 Future Research

A Critique of the Present Study

Different news events are known to have different impacts. Typically quarterly filings of reports and news of mergers and acquisitions are known to impact the markets on a different scale compared to news events such as product recalls or employee cutbacks. The relative importance of different news categories can be a research study in its own right. An obvious approach would be to apply the method of **event study** to determine the relative importance of these categories. The results can then be used to determine weightings of such news categories.

In our study we have used the bid-ask spread as the proxy for liquidity. Alternative liquidity measures, such as those mentioned in section 2.3, can be selected to carry out more modelling experiments. For example, the common depth measure of Kyle's λ can be used to explore the aspect of market depth in liquidity. Furthermore, the distribution assumption for liquidity can be changed to improve its multivariate predictions.

Other Research Directions

In our study we have taken a very general perspective whereby we find that the introduction of two different information streams (two different time series data), namely, news metadata and market data, improves the predictive power of asset behaviour. With this in mind, we see two further directions of research which can be of value to the financial modelling community.

It is well-known by practitioners that order flow data provides a good insight into trading strategies. Therefore, incorporation of order flow data as another time series is likely to improve the results of predictive models. We also observe that due to the growth of electronic communication networks (ECNs), such streams of data can be easily collected.

There has been growth in online posting of micro-blogs in general and Twitter data, to be more specific. As in the case of newswire texts which are transformed into sentiment data streams, Twitter data is also similarly processed (see Appendix C) and hence can be incorporated as an additional time series in the predictive models. We may also consider another source of online information, namely, search engine queries sequenced over time. For instance, if we consider *Google Trends*, it is possible to find a further information stream which can be used as yet another time series. The inclusion of such additional information streams is likely to give better predictions of asset behaviour.

References

1. Acharya, V. V. and Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics.* Vol. 77, No. 2, pp. 375–410.
2. AFM (2010). *High frequency trading: The application of advanced trading technology in the European marketplace.* Netherlands.
3. Aitken, M. and Comerton-Forde, C. (2003). How should liquidity be measured? *Pacific-Basin Finance Journal*. Vol. 11, pp. 45–59.
4. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. Vol. 19, No. 6, pp. 716–723.
5. Amihud, Y. and Mendelson, H. (1986). Asset Pricing and the Bid-Ask Spread. *Journal of Financial Econometrics.* Vol. 17, pp. 223–249.
6. Amihud, Y. (2002). Illiquidity and Stock returns: cross-section and time-series effects. *Journal of Financial Markets*. Vol. 5, pp. 31–56.
7. Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*. Vol. 59, No. 3, pp. 1259-1294.
8. Arbex-Valle, C., Erlwein-Sayer, C., Kochendörfer, A., Kübler, B., Mitra, G., Nzouankeu Nana, G. A., Nouwt, B. and Stalknecht, B. (2013). News-Enhanced Market Risk Management. Available at SSRN: http://ssrn.com/abstract=2322668
9. Arnuk, S. L. and Saluzzi, J (2008). *Toxic equity trading order flow on Wall Street: the real force behind the explosion in volume and volatility*. Available at: http://www.themistrading.com/article_files/0000/0524/Toxic_Equity_Trading_on_Wall_Street_--_FINAL_2__12.17.08.pdf
10. Bagehot, W. (1971). The only game in town. *Financial Analysts Journal*. Vol. 27, No. 2, pp. 12-14.
11. Baker, M. and Stein, J. C. (2004). Market liquidity as a sentiment indicator. *Journal of Financial Markets*. Vol. 7, No. 3, pp. 271-299.
12. Baker, M. and Wurgler, J. (2000). The equity share in new issues and aggregate stock returns. *The Journal of Finance.* Vol. 55, No. 5, pp. 2219-2257.
13. Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*. Vol. 61, No. 4, pp. 1645-1680.

14. Bandopadhyaya, A. and Jones, A. L. (2006). Measuring investor sentiment in equity markets. *Journal of Asset Management*. Vol. 7, No. 3, pp. 208-215.
15. Barber, B. M., and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behaviour of individual and institutional investors. *Review of Financial Studies*. Vol. 21, No. 2, pp. 785-818.
16. Barberis, N. (2000). Investing for the long run when returns are predictable. *The Journal of Finance*. Vol. 55, No. 1, pp. 225-264.
17. Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance.* Vol. 32, No. 3, pp. 663–682.
18. Bauwens, L., Laurent, S., and Rombouts, J. V. (2006) Multivariate GARCH models: a survey. *Journal of Applied Econometrics.* Vol. 21, No. 1, pp. 79-109.
19. Becker, K. G., Finnerty, J. E. and Kopecky, K. J. (1996) Macroeconomic news and the efficiency of international bond futures markets. *Journal of Futures Markets.* Vol. 16, No. 2, pp. 131-145.
20. Bernanke, B. S. and Kuttner, K. N. (2005). What explains the stock market's reaction to Federal Reserve policy? *The Journal of Finance*. Vol. 60, No. 3, pp. 1221-1257.
21. Berry, T. D., & Howe, K. M. (1994). Public information arrival. *The Journal of Finance*. Vol. 49, No. 4, pp. 1331-1346.
22. Black, F. (1971). Towards a fully automated exchange, part I. *Financial Analysts Journal.* Vol. 27, pp. 29–34.
23. Black, F. and Litterman, R. B. (1991). Asset allocation: combining investor views with market equilibrium. *The Journal of Fixed Income*. Vol. 1, No. 2, pp. 7-18.
24. Bloomfield, R., O’hara, M. and Saar, G. (2005) The “make or take” decision in an electronic market: Evidence on the evolution of liquidity. *Journal of Financial Economics*. Vol. 75, No. 1, pp. 165-199.
25. Bollen, J., Mao, H. and Zeng, X. (2011) Twitter mood predicts the stock market. *Journal of Computational Science.* Vol. 2, No. 1, pp. 1-8.
26. Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics.* Vol. 31, No. 3, pp. 307-327.
27. Bollerslev, T., Engle, R. F. and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *The Journal of Political Economy*, pp. 116-131.

28. Brennan, M. J. and Subrahmanyam, A. (1996) Market microstructure and asset pricing: On the compensation for illiquidity in stock returns. *Journal of financial economics*. Vol. 41, No. 3, pp. 441-464.

29. Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Vol. 60, No. 3, pp. 627-641.

30. Cahan, R., Jussa, J., and Luo, Y. (2009) Breaking news: How to use news sentiment to pick stocks. MacQuarie Research Report.

31. Campbell, J. Y., Grossman, S. J. and Wang, J. (1993) Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*. Vol. 108, No. 4, pp. 905-939.

32. Cappiello, L., Engle, R. and Sheppard, K. (2006) Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics.* Vol. 4, No. 4, pp. 537-572.

33. Chaboud, A., Chiquoine, B., Hjalmarsson, E. and Vega, C. (2013) Rise of the machines: Algorithmic trading in the foreign exchange market. *Journal of Finance, Forthcoming.*

34. Chan, W. S. (2003) Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*. Vol. 70, No. 2, pp. 223-260.

35. Chatfield, C. (2009). *The Analysis of Time Series: An Introduction.* 6th edition. Chapman & Hall Texts in Statistical Science.

36. Chen, C. W., Chiang, T. C. and So, M. K. (2003). Asymmetrical reaction to US stock-return news: evidence from major stock markets based on a double-threshold model. *Journal of Economics and Business*. Vol. 55, No. 5, pp. 487-502.

37. Chordia, T., Roll, R. and Subrahmanyam, A. (2000). Commonality in liquidity. *Journal of Financial Economics*. Vol. 56, No. 1, pp. 3-28.

38. Chordia, T., Roll, R. and Subrahmanyam, A. (2001). Market liquidity and trading activity. *The Journal of Finance*. Vol. 56, No. 2, pp. 501-530.

39. Cutler, D. M., Poterba, J. M. and Summers, L. H. (1989). What moves stock prices? *The Journal of Portfolio Management.* Vol. 15, No. 3, pp. 4-12.

40. Das, S. (2010). The finance web: internet information and markets. *IEEE Intelligent Systems.* Vol. 25, No. 2, pp. 74-78.

41. Das, S.Y. and Chen, M.Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science.* Vol. 53, No. 9, pp. 1375–1388.
42. Das, S. and Sisk, J. (2003). Financial communities. Available at SSRN: http://ssrn.com/abstract=404621
43. Davis, A. K., Piger, J. M. and Sedor, L. M. (2006). Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Federal Reserve Bank of St. Louis, Working paper Series,* (2006-005).
44. DeLong, J. B., Shleifer, A., Summers, L. H. and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy.* Vol. 98, pp. 703–738.
45. Dennis, P. and Mayhew, S. (2002). Risk-neutral skewness: Evidence from stock options. *Journal of Financial and Quantitative Analysis.* Vol. 37, No. 3, pp. 471-493.
46. Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association.* Vol. 74, *No.* 366, pp. 427–431.
47. Dion, M. (2013). Language recognition and news flow. Proceedings of *Behavioural Models and Sentiment Analysis Applied to Finance* in London 2 July 2013. Available at: http://unicom.co.uk/quant-finance/index.php.
48. Dzielinski, M. (2011). News sensitivity and the cross-section of stock returns. Available at SSRN: http://ssrn.com/abstract=1889030.
49. Dzielinski, M., Rieger, M. O. and Talpsepp, T. (2011). Volatility asymmetry, news and private investors. In Mitra, G. and Mitra, L. (ed.) *The Handbook of News Analytics in Finance.* John Wiley & Sons, pp. 255-270.
50. Easley, D. and O'hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial economics*. Vol. 19, No. 1, pp. 69-90.
51. Easley, D., Kiefer, N. M., O'hara, M. and Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *The Journal of Finance*. Vol. 51, No. 4, pp. 1405-1436.
52. Ederington, L. H. and Lee, J. H. (1993). How markets process information: News releases and volatility. *The Journal of Finance*. Vol. 48, No. 4, pp. 1161-1191.
53. Ederington, L. H. and Lee, J. H. (1995). The short-run dynamics of the price adjustment to new information. *Journal of Financial and Quantitative Analysis*. Vol. 30, No. 1, pp. 117-134.

54. Engelberg, J. E., Reed, A. V. and Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*. Vol. 105, No. 2, pp. 260-278.

55. Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*. Vol. 50, No. 4, pp. 987-1007.

56. Engle, R. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The journal of finance*. Vol. 48, No. 5, pp. 1749-1778.

57. Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*. Vol. 20, No. 3, pp. 339-350.

58. Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*. Vol. 47, No. 2, pp. 427-465.

59. Fioruci, J. A., Ehlers, R. S. and Andrade Filho, M. G. (2014). Bayesian multivariate GARCH models with dynamic correlations and asymmetric error distributions. *Journal of Applied Statistics*. Vol. 41, No. 2, pp. 320-331.

60. Fisher, K. L. and Statman, M. (2000) Investor sentiment and stock returns. *Financial Analysts Journal*. pp. 16-23.

61. Francioni, R., Hazarika, S., Reck, M. and Schwartz, R.A. (2008). Equity market microstructure: taking stock of what we know. *The Journal of Portfolio Management.* Vol. 35, No. 1, pp. 57-71.

62. Garman, M. B. (1976). Market microstructure. *Journal of financial Economics*. Vol. 3, No. 3, pp. 257-275.

63. Gidófalvi, G. and Elkan, C. (2001). Using news articles to predict stock price movements. *Technical Report. Department of Computer Science and Engineering, University of California, San Diego*.

64. Gomber, P., Arndt, B., Lutat, M. and Uhle, T. (2011). High frequency trading. Available at SSRN: http://ssrn.com/abstract=1858626.

65. Goodhart, C. A. and O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance.* Vol. 4, No.2, pp. 73-114.

66. Gross-Klussmann, A. and Hautsch, N (2011). When machines read the news: using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance.* Vol. 18, pp. 321–340.

67. Hafez, P. A. and Xie, J. (2012). Factoring sentiment risk into quant models. Available at SSRN: http://ssrn.com/abstract=2071142.

68. Hafez, P. (2013). Market-level Sentiment for trading Forex and equity indices. Proceedings of *Behavioural Models and Sentiment Analysis Applied to Finance* in London 2 July 2013. Available at: http://unicom.co.uk/quant-finance/index.php.
69. Harris, L. (1998). Optimal dynamic order submission strategies in some stylized trading problems. *Financial Markets, Institutions & Instruments*. Vol. 7, No. 2, pp. 1-76.
70. Harris, L. (2002) *Trading and Exchanges: Market Microstructure for Practitioners*. New York: Oxford University Press.
71. Harris, L. (2013). What to do about high frequency trading? *Financial Analysts Journal*. Vol. 69, No. 2, pp. 6-9.
72. Harvey, A. C. (1990). *Forecasting structural time series models and the Kalman filter*. Cambridge University Press.
73. Hasbrouck, J. (2006). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. New York: Oxford University Press.
74. Hasbrouck, J. and Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of financial Economics*. Vol. 59, No. 3, pp. 383-411.
75. Hasbrouck, J. and Schwartz, R. A. (1988). Liquidity and execution cost in equity markets. *The Journal of Portfolio Management*. Vol. 14, pp. 10–16.
76. Ho, K. Y., Shi, Y. and Zhang, Z. (2013). How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches. *The North American Journal of Economics and Finance*. Vol. 26, pp. 436-456.
77. Hong, H. and Rady, S. (2002). Strategic trading and learning about liquidity. *Journal of Financial Markets*. Vol. 5, No. 4, pp. 419-450.
78. Hui, B. and Heubel, B. (1984) *Comparative liquidity advantages among major U.S. stock markets*. Data Resources inc.
79. Jacquier, E. and Polson, N. (2011). Bayesian Methods in Finance. In Geweke, J., Koop, G. and van Dijk, H. (eds) *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press. pp. 439-512
80. Kahn, R. (2013). Quant 3.0: Harnessing the mood of the web in alpha strategies. Proceedings of *Behavioural Models and Sentiment Analysis Applied to Finance* in London 2 July 2013. Available at: http://unicom.co.uk/quant-finance/index.php.

81. Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica.* Vol. 47, No. 2, pp. 263 – 292.
82. Kahneman, D. (2002). Maps of bounded rationality: The [2002] Sveriges Riksbank Prize. *[Lecture] in Economic Sciences.* Available at*:* http://nobelprize.org/nobel_prizes/.../2002/kahneman-lecture.html
83. Kearns, M., Kulesza, A. and Nevmyvaka, Y. (2010). Empirical limitations on high frequency trading profitability. Available at SSRN*:* http://ssrn.com/abstract=1678758.
84. Kim, O. and Verrecchia, R. E. (1994). Market liquidity and volume around earnings announcements. *Journal of Accounting and Economics.* Vol. 17, No. 1-2, pp. 41-67.
85. Kothari, S. P. and Shanken, J. (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics*. Vol. 44, No. 2, pp. 169-203.
86. Krinsky, I. and Lee, J. (1996). Earnings announcements and the components of the bid-ask spread. *The Journal of Finance.* Vol. 51, No. 4, pp. 1523-1535.
87. Kumar, M. S. and Persaud, A. (2002). Pure contagion and investors' shifting risk appetite: analytical issues and empirical evidence. *International Finance*. Vol. 5, No. 3, pp. 401-436.
88. Kyle, A. (1985) Continuous auction and insider trading. *Econometrica*. Vol. 53, pp. 1315–35.
89. Lashgari, M. (2000). The role of TED spread and confidence index in explaining the behavior of stock prices. *American Business Review*. Vol. 18, No. 2, pp. 9-11.
90. LeBaron, B., Arthur, W. B. and Palmer, R. (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamics and control*. Vol. 23, No. 9, pp. 1487-1516.
91. Lee, C.M.C. (1992). Earnings News and Small Traders. *Journal of Accounting and Economics.* Vol. 15, pp. 265-302.
92. Leinweber, D. (2009). *Nerds on Wall Street.* New Jersey: John Wiley & Sons.
93. Leinweber, D. and Sisk, J. (2011). Relating news analytics to stock returns. In Mitra, G. and Mitra, L. (ed.) *The Handbook of News Analytics in Finance*. John Wiley & Sons, pp. 149-172.
94. Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? Available at SSRN: http://ssrn.com/abstract=898181.

95. Li, L. and Engle, R. F. (1998). Macroeconomic Announcements and Volatility of Treasury Futures. *UCSD Economics Discussion Paper 98-27.* Available at SSRN: http://ssrn.com/abstract=145828

96. Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in the portfolios and capital budgets. *Review of Economics and Statistics.* Vol. 47, pp. 13-37.

97. Litterman, R. (1980). A Bayesian procedure for forecasting with vector autoregressions. *Mimeo, Massachussets Institute of Technology.*

98. Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*. Vol. 66, No. 1, pp. 35-65.

99. Lütkepohl, H. (1993). *Introduction to multiple time series analysis.* New York: Springer.

100. Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets*. Vol. 3, No. 3, pp. 205-258.

101. Mahler, N. (2009). Modelling the S&P 500 index using the Kalman filter and the LagLasso. In *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on* (pp. 1-6). IEEE.

102. Malek, L. (2000). The role of TED spread and confidence index in explaining the behaviour of stock prices. *American Business Review*. Vol. 18, No. 2, pp. 9-11.

103. Mitra, G. and Mitra, L. (2011). *The Handbook of News Analytics in Finance.* John Wiley & Sons.

104. Mitra, L., Mitra, G., and diBartolomeo, D., (2009). Equity portfolio risk (volatility) estimation using market information and sentiment. *Quantitative Finance.* Vol. 9, No. 8, pp. 887-895.

105. Moniz, A., Brar, G. and Davis, C. (2009). *Have I got news for you*. MacQuarie Research Report.

106. Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the Econometric Society*. pp. 768-783.

107. Neal, R. and Wheatley, S. M. (1998). Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis*. Vol. 33, No. 4, pp. 523-547.

108. Nicholson, S. F. (1968). Price ratios in relation to investment results. *Financial Analysts Journal.* pp. 105–109.

109. Niederhoffer, V. (1971). The analysis of world events and stock prices. *The Journal of Business*. Vol. 44, No. 2, pp. 193-219.
110. Nocera, J. (2009). "*Poking Holes in a Theory on Markets*". New York Times.
111. O'hara, M. (1995). *Market microstructure theory* (Vol. 108). Cambridge, MA: Blackwell.
112. Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of *ACL-02 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, Philadelphia, PA: Volume 10 (pp. 79-86).
113. Patell, J. M. and Wolfson, M. A. (1984). The intraday speed of adjustment of stock prices to earnings and dividend announcements. *Journal of Financial Economics*. Vol. 13, No. 2, pp. 223-252.
114. Patton, A. J. and Verardo, M. (2012). Does beta move with news? Firm-specific information flows and learning about profitability. *Review of Financial Studies.* Vol. 25, No. 9, pp. 2789-2839.
115. Peterson, R.L. (2007). *Inside the Investor's Brain*. New Jersey: John Wiley & Sons.
116. Preis, T., Moat, H. S. and Stanley, H. E. (2013). Quantifying trading behaviour in financial markets using Google Trend. *Scientific Reports*, *3.*
117. Rachev, S. T., Hsu, J. S. J., Bagasheva, B. S. and Fabozzi, F. J. (2008). *Bayesian Methods in Finance.* New Jersey: John Wiley & Sons.
118. Ranaldo, A. (2008). Intraday market dynamics around public information arrivals. In Lhabitant, F.S. and Gregoriou G. N. (ed.) *Stock market liquidity: Implications for market microstructure and asset pricing*. New Jersey: John Wiley & Sons, pp. 199-226.
119. Randall, M. R., Suk, D. Y. and Tully, S. W. (2003). Mutual Fund Cash Flows and Stock Market Performance. *The Journal of Investing*. Vol. 12, No. 1, pp. 78-80.
120. RavenPack (2010). RavenPack News Scores: Forward-looking News Analysis. User Guide to Data and Service Overview. Version 1.5.1.
121. Reinsel, G. C. (2003). *Elements of multivariate time series analysis*. 2nd edition. New York: Springer.
122. Rigobon, R. and Sack, B. (2004). The impact of monetary policy on asset prices. *Journal of Monetary Economics*. Vol. 51, No. 8, pp. 1553-1575.

123. Riordan, R., Storkenmaier, A., Wagener, M. and Zhang, S. (2013). Public information arrival: Price discovery and liquidity in electronic limit order markets. *Journal of Banking & Finance.* Vol. 37, No. 4, pp. 1148-1159.

124. Robertson, D. and Wright, S. (2009). *The Limits to Stock Return Predictability*. Available at: http://www.ems.bbk.ac.uk/faculty/wright/pdf/limits.

125. Rosenberg, B., Reid, K. and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management.* Vol. 11, No. 3, pp. 9–16.

126. Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial management review*. Vol. 6, No. 2, pp. 41-49.

127. Schumaker, R. P., Zhang, Y., Huang, C. and Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems.* Vol. 53, pp. 458-464.

128. Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics.* Vol. 6, No. 2, pp. 461–464.

129. Seasholes, M. and Wu, G. (2004). Profiting from predictability: Smart traders, daily price limits, and investor attention. *University of California, Berkeley, working paper*.

130. Simon, H. A. (1964). On the concept of organizational goal. *Administrative Science Quarterly*. pp. 1-22.

131. Sinha, N. (2010). Underreaction to news in the US stock market. Available at SSRN: http://ssrn.com/abstract=1572614.

132. Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of finance*. Vol. 19, No. 3, pp. 425-442.

133. Shefrin, H. (2008). *A Behavioral Approach to Asset Pricing*. Elsevier.

134. Shiller, R. J., Fischer, S. and Friedman, B. M. (1984). Stock prices and social dynamics. *Brookings Papers on Economic Activity*. Vol. 2, pp. 457-510.

135. Shiller, R. (2000). *Irrational Exuberance*. Princeton University Press.

136. Smales, L. A. (2012). Non-scheduled news arrival and high-frequency stock market dynamics: Evidence from the Australian Securities Exchange. *25th Australasian Finance and Banking Conference 2012.* Available at SSRN: http://ssrn.com/abstract=2130193.

137. Smales, L. A. (2013). News sentiment and the investor fear gauge. *Finance Research Letters*. Available at: http://dx.doi.org/10.1016/j.frl.2013.07.003

138. Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*. Vol. 54, No. 3, pp. 375-421.

139. Talsepp, T. and Reiger, M. (2009). Explaining asymmetric volatility around the world. *Journal of Empirical Finance.* Vol. 17, No. 5, pp. 938-956.

140. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*. Vol. 62, No. 3, pp. 1139-1168.

141. Tetlock, P. C., Saar-Tsechansky, M.A.Y.T.A.L. and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*. Vol. 63, No. 3, pp. 1437-1467.

142. Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies*. Vol. 24, No. 5, pp.1481-1512.

143. Thomson Reuters (2010). Thomson Reuters News Analytics. Version 2.0.2.

144. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. pp. 267-288.

145. Treynor, J. L. (1961). Toward a theory of market value of risky assets. Unpublished manuscript. A final version was published in *Asset Pricing and Portfolio Performance* (2009).

146. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the *40th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics (ACL'02), Philadelphia, PA: pp. 417-424.

147. Uhl, M. W. (2011). Reuters sentiment and stock returns (No. 288). KOF working papers/KOF Swiss Economic Institute, ETH Zurich.

148. Vrontos, I. D., Dellaportas, P. and Politis, D. N. (2003). A full-factor multivariate GARCH model. *The Econometrics Journal*. Vol. 6, No. 2, pp. 312-334.

149. Whaley, R. (2000). The investor fear gauge. *The Journal of Portfolio Management.* Vol. 26, No. 3, pp. 12-17.

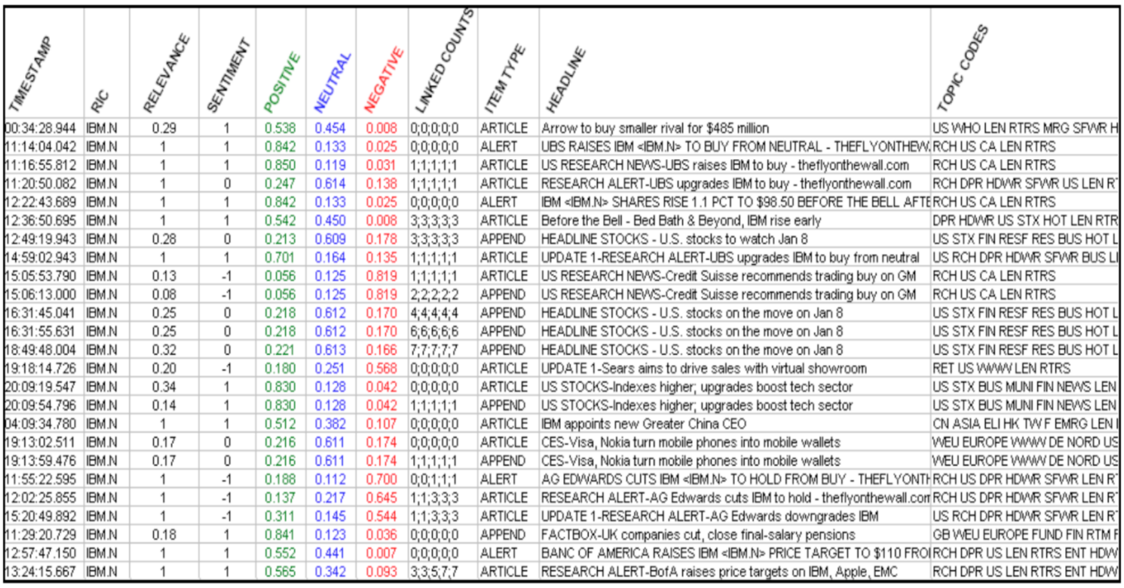
150. Woodruff, C. S. and Senchack, A. J. (1988). Intradaily price-volume adjustments of NYSE stocks to unexpected earnings. *The Journal of Finance*. Vol. 43, No. 2, pp. 467-491.

151. Zhang, X., Fuehres, H. and Gloor, P. A. (2011). Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. *Procedia-Social and Behavioral Sciences*. Vol. 26, pp. 55-62.

Appendix A

News Sentiment Metadata from Thomson Reuters

We provide a snapshot of the news sentiment metadata produced by Thomson Reuters in a tabular form. Here the sample company is chosen to be IBM and news events are listed for a few consecutive days. Not all data attributes are displayed.



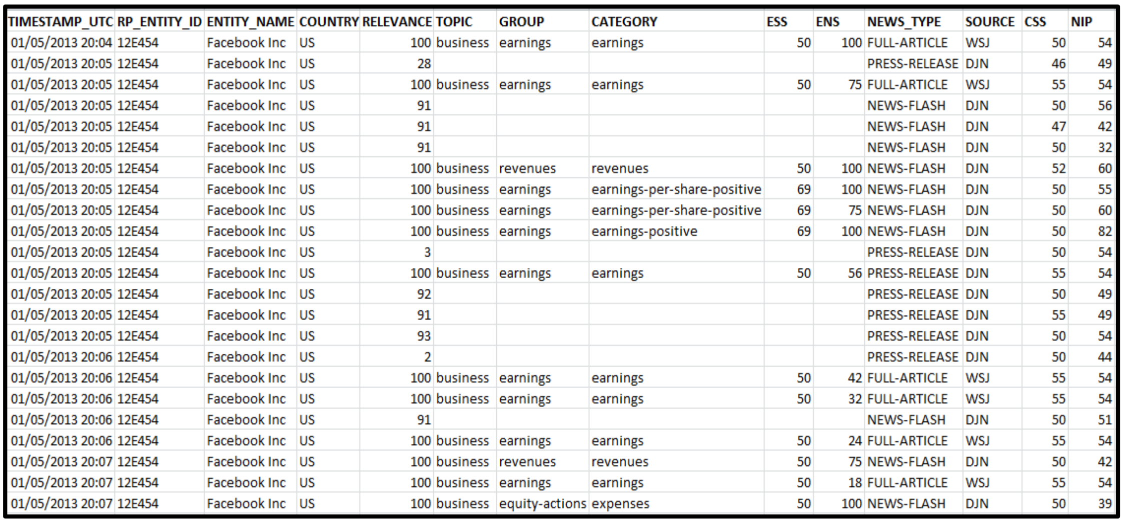
TIMESTAMP	RIC	RELEVANCE	SENTIMENT	POSITIVE	NEUTRAL	NEGATIVE	LINKED COUNTS	ITEM TYPE	HEADLINE	TOPIC CODES
00:34:28.944	IBM.N	0.29	1	0.538	0.454	0.008	0;0;0;0;0	ARTICLE	Arrow to buy smaller rival for $485 million	US WHO LEN RTRS MRG SFWR H
11:14:04.042	IBM.N	1	1	0.842	0.133	0.025	0;0;0;0;0	ALERT	UBS RAISES IBM <IBM.N> TO BUY FROM NEUTRAL - THEFLYONTHEW.	RCH US CA LEN RTRS
11:16:55.812	IBM.N	1	1	0.850	0.119	0.031	1;1;1;1;1	ARTICLE	US RESEARCH NEWS-UBS raises IBM to buy - theflyonthewall.com	RCH US CA LEN RTRS
11:20:50.082	IBM.N	1	0	0.247	0.614	0.138	1;1;1;1;1	ARTICLE	RESEARCH ALERT-UBS upgrades IBM to buy - theflyonthewall.com	RCH DPR HDWR SFWR US LEN R
12:22:43.689	IBM.N	1	1	0.842	0.133	0.025	0;0;0;0;0	ALERT	IBM <IBM.N> SHARES RISE 1.1 PCT TO $98.50 BEFORE THE BELL AFTE	RCH US CA LEN RTRS
12:36:50.695	IBM.N	1	1	0.542	0.450	0.008	3;3;3;3;3	ARTICLE	Before the Bell - Bed Bath & Beyond, IBM rise early	DPR HDWR US STX HOT LEN RTR
12:49:19.943	IBM.N	0.28	0	0.213	0.609	0.178	3;3;3;3;3	APPEND	HEADLINE STOCKS - U.S. stocks to watch Jan 8	US STX FIN RESF RES BUS HOT L
14:59:02.943	IBM.N	1	1	0.701	0.164	0.135	1;1;1;1;1	ARTICLE	UPDATE 1-RESEARCH ALERT-UBS upgrades IBM to buy from neutral	US RCH DPR HDWR SFWR BUS LI
15:05:53.790	IBM.N	0.13	-1	0.056	0.125	0.819	1;1;1;1;1	ARTICLE	US RESEARCH NEWS-Credit Suisse recommends trading buy on GM	RCH US CA LEN RTRS
15:06:13.000	IBM.N	0.08	-1	0.056	0.125	0.819	2;2;2;2;2	APPEND	US RESEARCH NEWS-Credit Suisse recommends trading buy on GM	RCH US CA LEN RTRS
16:31:45.041	IBM.N	0.25	0	0.218	0.612	0.170	4;4;4;4;4	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
16:31:55.631	IBM.N	0.25	0	0.218	0.612	0.170	6;6;6;6;6	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
18:49:48.004	IBM.N	0.32	0	0.221	0.613	0.166	7;7;7;7;7	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
19:18:14.726	IBM.N	0.20	-1	0.180	0.251	0.568	0;0;0;0;0	ARTICLE	UPDATE 1-Sears aims to drive sales with virtual showroom	RET US WWW LEN RTRS
20:09:19.547	IBM.N	0.34	1	0.830	0.128	0.042	0;0;0;0;0	ARTICLE	US STOCKS-Indexes higher; upgrades boost tech sector	US STX BUS MUNI FIN NEWS LEN
20:09:54.796	IBM.N	0.14	1	0.830	0.128	0.042	1;1;1;1;1	APPEND	US STOCKS-Indexes higher; upgrades boost tech sector	US STX BUS MUNI FIN NEWS LEN
04:09:34.780	IBM.N	1	1	0.512	0.382	0.107	0;0;0;0;0	ARTICLE	IBM appoints new Greater China CEO	CN ASIA ELI HK TW F EMRG LEN I
19:13:02.511	IBM.N	0.17	0	0.216	0.611	0.174	0;0;0;0;0	ARTICLE	CES-Visa, Nokia turn mobile phones into mobile wallets	WEU EUROPE WWW DE NORD US
19:13:59.476	IBM.N	0.17	0	0.216	0.611	0.174	1;1;1;1;1	APPEND	CES-Visa, Nokia turn mobile phones into mobile wallets	WEU EUROPE WWW DE NORD US
11:55:22.595	IBM.N	1	-1	0.188	0.112	0.700	0;0;1;1;1	ALERT	AG EDWARDS CUTS IBM <IBM.N> TO HOLD FROM BUY - THEFLYONTH	RCH US DPR HDWR SFWR LEN R
12:02:25.855	IBM.N	1	-1	0.137	0.217	0.645	1;1;3;3;3	ARTICLE	RESEARCH ALERT-AG Edwards cuts IBM to hold - theflyonthewall.con	RCH US DPR HDWR SFWR LEN R
15:20:49.892	IBM.N	1	-1	0.311	0.145	0.544	1;1;3;3;3	ARTICLE	UPDATE 1-RESEARCH ALERT-AG Edwards downgrades IBM	US RCH DPR HDWR SFWR LEN R
11:29:20.729	IBM.N	0.18	1	0.841	0.123	0.036	0;0;0;0;0	APPEND	FACTBOX-UK companies cut, close final-salary pensions	GB WEU EUROPE FUND FIN RTM F
12:57:47.150	IBM.N	1	1	0.552	0.441	0.007	0;0;0;0;0	ALERT	BANC OF AMERICA RAISES IBM <IBM.N> PRICE TARGET TO $110 FRO	RCH DPR US LEN RTRS ENT HDW
13:24:15.667	IBM.N	1	1	0.565	0.342	0.093	3;3;5;7;7	ARTICLE	RESEARCH ALERT-BofA raises price targets on IBM, Apple, EMC	RCH DPR US LEN RTRS ENT HDW

Figure A: Sample of news sentiment data. *Source: Thomson Reuters*. Further details are given in Thomson Reuters (2010).

Appendix B

News Sentiment Metadata from RavenPack

We display below a sample of news sentiment metadata produced by RavenPack in a tabular form. The most important data fields have been selected and included in the snapshot for news regarding Facebook Inc.



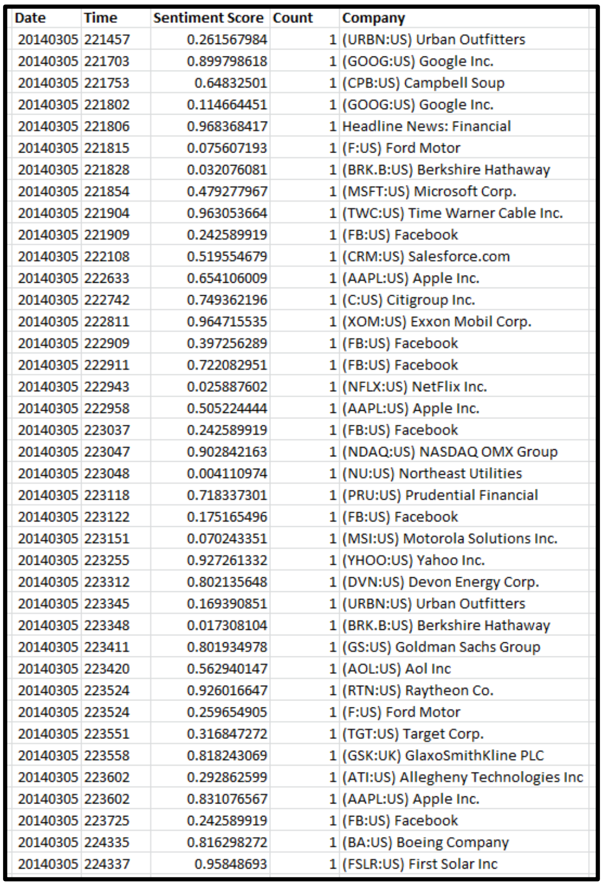
TIMESTAMP_UTC	RP_ENTITY_ID	ENTITY_NAME	COUNTRY	RELEVANCE	TOPIC	GROUP	CATEGORY	ESS	ENS	NEWS_TYPE	SOURCE	CSS	NIP
01/05/2013 20:04	12E454	Facebook Inc	US	100	business	earnings	earnings	50	100	FULL-ARTICLE	WSJ	50	54
01/05/2013 20:05	12E454	Facebook Inc	US	28						PRESS-RELEASE	DJN	46	49
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	earnings	earnings	50	75	FULL-ARTICLE	WSJ	55	54
01/05/2013 20:05	12E454	Facebook Inc	US	91						NEWS-FLASH	DJN	50	56
01/05/2013 20:05	12E454	Facebook Inc	US	91						NEWS-FLASH	DJN	47	42
01/05/2013 20:05	12E454	Facebook Inc	US	91						NEWS-FLASH	DJN	50	32
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	revenues	revenues	50	100	NEWS-FLASH	DJN	52	60
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	earnings	earnings-per-share-positive	69	100	NEWS-FLASH	DJN	50	55
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	earnings	earnings-per-share-positive	69	75	NEWS-FLASH	DJN	50	60
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	earnings	earnings-positive	69	100	NEWS-FLASH	DJN	50	82
01/05/2013 20:05	12E454	Facebook Inc	US	3						PRESS-RELEASE	DJN	50	54
01/05/2013 20:05	12E454	Facebook Inc	US	100	business	earnings	earnings	50	56	PRESS-RELEASE	DJN	55	54
01/05/2013 20:05	12E454	Facebook Inc	US	92						PRESS-RELEASE	DJN	50	49
01/05/2013 20:05	12E454	Facebook Inc	US	91						PRESS-RELEASE	DJN	55	49
01/05/2013 20:05	12E454	Facebook Inc	US	93						PRESS-RELEASE	DJN	50	54
01/05/2013 20:06	12E454	Facebook Inc	US	2						PRESS-RELEASE	DJN	50	44
01/05/2013 20:06	12E454	Facebook Inc	US	100	business	earnings	earnings	50	42	FULL-ARTICLE	WSJ	55	54
01/05/2013 20:06	12E454	Facebook Inc	US	100	business	earnings	earnings	50	32	FULL-ARTICLE	WSJ	55	54
01/05/2013 20:06	12E454	Facebook Inc	US	91						NEWS-FLASH	DJN	50	51
01/05/2013 20:06	12E454	Facebook Inc	US	100	business	earnings	earnings	50	24	FULL-ARTICLE	WSJ	55	54
01/05/2013 20:07	12E454	Facebook Inc	US	100	business	revenues	revenues	50	75	NEWS-FLASH	DJN	50	42
01/05/2013 20:07	12E454	Facebook Inc	US	100	business	earnings	earnings	50	18	FULL-ARTICLE	WSJ	55	54
01/05/2013 20:07	12E454	Facebook Inc	US	100	business	equity-actions	expenses	50	100	NEWS-FLASH	DJN	50	39

Figure B: Sample of news sentiment data. *Source: RavenPack.* Further details are available in RavenPack (2010).

Appendix C

Twitter Data in a Comparable Form

Twitter, a micro-blogging platform, now allows users to download content from the site. Therefore, a similar process of filtering and aggregation can be applied to this data in order to deduce the sentiment of the text. Figure C shows an example of such an information stream where sentiment probabilities have been assigned to each individual blog post (otherwise known as ‘Tweet’).



Date	Time	Sentiment Score	Count	Company
20140305	221457	0.261567984	1	(URBN:US) Urban Outfitters
20140305	221703	0.899798618	1	(GOOG:US) Google Inc.
20140305	221753	0.64832501	1	(CPB:US) Campbell Soup
20140305	221802	0.114664451	1	(GOOG:US) Google Inc.
20140305	221806	0.968368417	1	Headline News: Financial
20140305	221815	0.075607193	1	(F:US) Ford Motor
20140305	221828	0.032076081	1	(BRK.B:US) Berkshire Hathaway
20140305	221854	0.479277967	1	(MSFT:US) Microsoft Corp.
20140305	221904	0.963053664	1	(TWC:US) Time Warner Cable Inc.
20140305	221909	0.242589919	1	(FB:US) Facebook
20140305	222108	0.519554679	1	(CRM:US) Salesforce.com
20140305	222633	0.654106009	1	(AAPL:US) Apple Inc.
20140305	222742	0.749362196	1	(C:US) Citigroup Inc.
20140305	222811	0.964715535	1	(XOM:US) Exxon Mobil Corp.
20140305	222909	0.397256289	1	(FB:US) Facebook
20140305	222911	0.722082951	1	(FB:US) Facebook
20140305	222943	0.025887602	1	(NFLX:US) NetFlix Inc.
20140305	222958	0.505224444	1	(AAPL:US) Apple Inc.
20140305	223037	0.242589919	1	(FB:US) Facebook
20140305	223047	0.902842163	1	(NDAQ:US) NASDAQ OMX Group
20140305	223048	0.004110974	1	(NU:US) Northeast Utilities
20140305	223118	0.718337301	1	(PRU:US) Prudential Financial
20140305	223122	0.175165496	1	(FB:US) Facebook
20140305	223151	0.070243351	1	(MSI:US) Motorola Solutions Inc.
20140305	223255	0.927261332	1	(YHOO:US) Yahoo Inc.
20140305	223312	0.802135648	1	(DVN:US) Devon Energy Corp.
20140305	223345	0.169390851	1	(URBN:US) Urban Outfitters
20140305	223348	0.017308104	1	(BRK.B:US) Berkshire Hathaway
20140305	223411	0.801934978	1	(GS:US) Goldman Sachs Group
20140305	223420	0.562940147	1	(AOL:US) Aol Inc
20140305	223524	0.926016647	1	(RTN:US) Raytheon Co.
20140305	223524	0.259654905	1	(F:US) Ford Motor
20140305	223551	0.316847272	1	(TGT:US) Target Corp.
20140305	223558	0.818243069	1	(GSK:UK) GlaxoSmithKline PLC
20140305	223602	0.292862599	1	(ATI:US) Allegheny Technologies Inc
20140305	223602	0.831076567	1	(AAPL:US) Apple Inc.
20140305	223725	0.242589919	1	(FB:US) Facebook
20140305	224335	0.816298272	1	(BA:US) Boeing Company
20140305	224337	0.95848693	1	(FSLR:US) First Solar Inc

Figure C: Sample of Twitter sentiment data. *Source: FSWire*. Further information can be found at www.fswire.com.