

Applications of news analytics in finance: A review

Leela Mitra ^{*†} Gautam Mitra ^{*†}

June 17, 2010

Contents

^{*}CARISMA (Centre for Analysis of Risk and Optimisation Modelling Applications), Brunel University, Uxbridge, United Kingdom, UB8 3PH

[†]OptiRisk Systems, OptiRisk R&D House, One Oxford Road, Uxbridge, Middlesex, UB9 4DA UNITED KINGDOM

Abstract

A review of news analytics and its applications in finance is given in this chapter. In particular we review the multiple facets of current research and some of the major applications. It is widely recognised news plays a key role in financial markets. The sources and volumes of news continue to grow. New technologies that enable automatic or semi-automatic news collection, extraction, aggregation and categorisation are emerging. Further machine learning techniques can be used to process the textual input of news stories to determine quantitative sentiment scores. We consider the various types of news available and how these can be processed to form inputs to financial models. We report applications of news, for prediction of abnormal returns, for trading strategies, for diagnostic applications as well as the use of news for risk control.

1 Introduction

General value of news analysis for the asset management process

News (North, East, West, South) streams in from all parts of the globe. There is a strong yet complex relationship between market sentiment and news. The arrival of news continually updates investor's understanding and knowledge of the market and influences investor sentiment. There is a growing body of research literature that argues media influences investor sentiment, hence asset prices, asset price volatility and risk (Tetlock 2007, Da, Engleberg and Gao 2009, Odean and Barber 2008, diBartolomeo and Warrick 2005, Mitra, Mitra and diBartolomeo 2009, Dzielinski, Rieger and Talpsepp 2010). Traders and other market participants digest news rapidly, revising and rebalancing their asset positions accordingly. Most traders have access to newswires at their desks. As markets react rapidly to news, effective models which incorporate news data are highly sought after. This is not only for trading and fund management, but also for risk control. Major news events can have a significant impact on the market environment and investor sentiment resulting in rapid changes to the risk structure and risk characteristics of traded assets. Though the relevance of news is widely acknowledged how to incorporate this effectively, in quantitative models and more generally within the investment decision making process, is a very open question.

In considering how news impacts markets, Odean and Barber (2008) note "significant news will often affect investors' beliefs and portfolio goals heterogeneously, resulting in more investors trading than is usual" (high trading volume). It is well known volume increases on days with information releases (Bamber, Barron & Stober 1997, Karpoff 1987, Busse & Green 2002). Important news frequently results in large positive or negative returns. Ryan & Taffler (2002) find for large firms a significant portion (65%) of large price changes and volume movements can be linked to publicly available news releases. Sometimes investors may find it difficult to interpret news resulting in high trading volume without significant price change.

Financial news can be split into regular synchronous announcements (expected news) and event driven asynchronous announcements (unexpected news). Textual news is frequently unstructured, qualitative data. It is characterised as being non-numeric and hard to quantify. Unlike analysis based on quantified market data textual news data contains information about the effect of an event and the possible causes of an event. It is natural to expect that the application of this news data will lead to improved analysis (such as, predictions of returns and volatility) . However, extracting this information in a form that can be applied to the investment decision making process is extremely challenging.

News has always been a key source of investment information. The volumes and sources of news are growing rapidly. In increasingly competitive markets investors and traders need to select and analyse the relevant news, from the vast amounts available to them, in order to make "good" and timely decisions. A human's (or even a group of humans) ability to process this news is limited. As computational capacity grows, technologies are emerging which allow us to extract, aggregate and categorise large volumes of news effectively. Such technology might be applied for quantitative model construction for both high frequency trading and low frequency fund rebalancing. Automated news analysis can form a key component driving algorithmic trading desks' strategies and execution, and the traders who use this technology can shorten the time it takes them to react to breaking stories (that is, reduce latency times). News analytics (NA) technology can also

be used to aid traditional non-quantitative fund managers in monitoring the market sentiment for particular stocks, companies, brands and sectors. These technologies are deployed to automate filtering, monitoring and aggregation of news. These technology aids free managers from the minutae of repetitive analysis, such that they are able to better target their reading and research. These technologies reduce the burden of the routine monitoring for fundamental managers.

The basic idea behind these NA technologies is to automate human thinking and reasoning. Traders, speculators and private investors anticipate the direction of asset returns as well as, the size and the level of uncertainty (volatility) before making an investment decision. They carefully read recent economic and financial news to gain a picture of current situation. Using their knowledge of how markets behaved in the past, under different situations, people will implicitly match the current situation with those situations in the past most similar to the current one. News analytics seeks to introduce technology to automate or semi-automate this approach. 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 (reducing latency). Automation or semi-automation of the human judgement process widens the limits of the investment process. Leinweber(2009) refers to this process as Intelligence Amplification (IA).

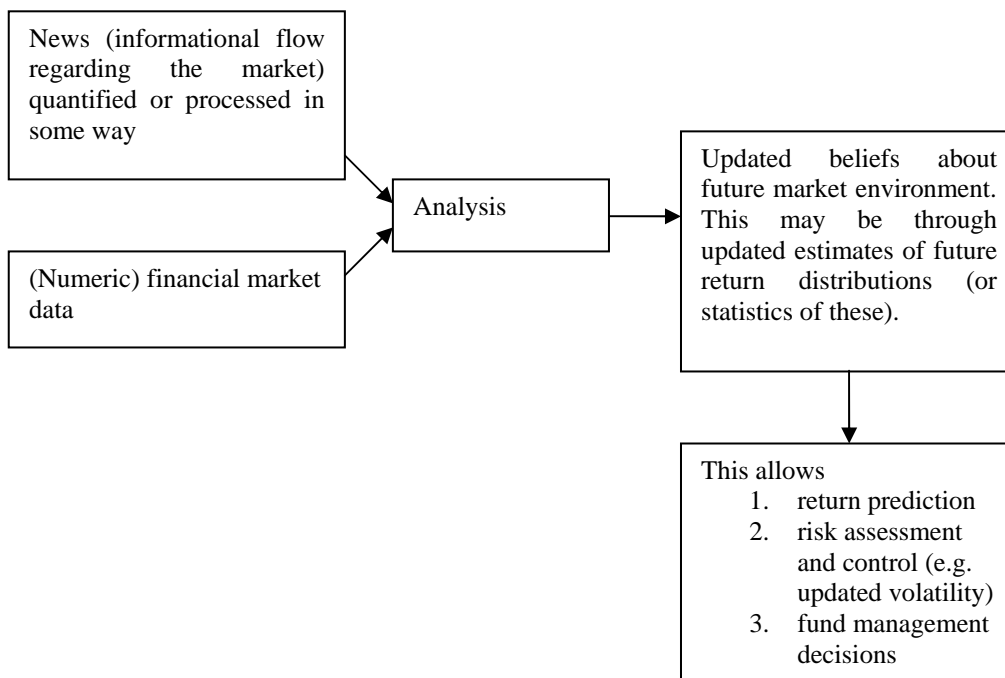


Figure 1: A simple representation of news analytics in financial decision making

As shown in Figure 1 news data is an additional source of information that can be harnessed to enhance (traditional) investment analysis. Yet it is important to recognise that NA in finance is a multi disciplinary field which draws on financial economics, financial engineering, behavioural finance and Artificial Intelligence (in particular Natural Language Processing). Expertise in these respective areas need to be combined effectively for the development of successful applications in this area. Sophisticated machine learning algorithms applied without an understanding of the structure and dynamics of financial markets and the use of realistic trading assumptions can lead to applications with little commercial use (See Mittermayer and Knolmayer 2006).

The remainder of the chapter is organised as follows. In section 2.1 we consider the different sources of

news and information flows which can be applied for updating (quantitative) investor beliefs and knowledge. Section 2.2 covers several aspects of pre-analysis to be considered when using news in trading systems and quantitative models. In section 3 we consider how qualitative text can be converted to quantified metrics which can form inputs to quantitative models. In section 4 we present news based models; In particular we consider the computational architecture (section 4.1), applications for trading and fund management (section 4.2) and applications for risk management (section 4.3). In section 4.4 desirable industry applications are outlined. The appendix ?? contains an annotated bibliography of selected papers.

2 News data

2.1 Data sources

In this section we consider the different sources of news and information flows which can be applied for updating (quantitative) investor beliefs and knowledge. Leinweber (2009) distinguishes four broad classifications of news (informational flows).

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 may also exist.
2. **Pre-News** This refers to the source data 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 macro economic news, industry statistics, company earnings reports and other corporate news.
3. **Rumours** These are blogs and websites that broadcast "news", and are less reputable than news and pre-news sources. The quality of these vary significantly. Some may be blogs associated with highly reputable news providers and reporters (for example, Robert Peston of the BBC's blog!). At the other end of the scale some blogs may lack any substance and may be entirely fueled by rumour.
4. **Social media** These websites fall at the lowest end of the reputation scale. Barriers to entry are extremely low and the ability to publish "information" easy. These can be dangerously inaccurate sources of information. However, if carefully applied (with consideration of human behaviour and agendas) there may be some value to be gleaned from these. At a minimum they may help us identify future volatility.

Individual investors pay more attention to the second two sources of news than institutional investors. (Dzielinski, Rieger and Talpsepp 2010 and Das & Chen 2007). 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 Ch 10 2009). The SEC does monitor message boards. So there is some, though perhaps far from perfect, checking of information published. This should constrain message board posters actions to some extent.

There are services which facilitate retrieval of news data from the web. For example the Google trends is a free but limited service which provides historical weekly time series of the popularity of any given search term. This search engine reports the proportion of positive, negative and neutral stories returned for a given search.

The Securities and Exchange Commission (SEC) provides a lot of useful pre news. It covers all publicly traded companies (in the US). The Electronic Data Gathering, Analysis and Retrieval (EDGAR) system was introduced in 1996 giving basic access to filings via the web. (See <http://www.sec.gov/edgar.shtml>) Premium access gave tools for analysis of filing information and priority earlier access to the data. In 2002 filing information was released to the public in real time. Filings remain unstructured text files without semantic Web and XML output, though the SEC are in the process of upgrading their information dissemination. High end resellers electronically dissect and sell on relevant component parts of filings. Managers are obliged to disclose a significant amount of information about a company via SEC filings. This information

is naturally valuable to investors. Leinweber introduces the term “molecular search: the idea of looking for patterns and changes in groups of documents”. Such analysis/information are scrutinized by researchers/analysts to identify unusual corporate activity and potential investment opportunities. However mining the large volume of filings, to find relationships, is challenging. Engleberg and Sankaraguruswamy (2007) note the EDGAR database has 605 different forms and there were 4,249,586 filings between 1994 and 2006. Connotate provides services which allows customised automated collection of SEC filing information for customers (fund managers and traders). Engleberg and Sankaraguruswamy (2007) consider how to use a SAS web crawler to mine SEC filing information through EDGAR.

As stated in section 1, financial news can be split into *regular synchronous announcements (scheduled or expected news)* and *event driven asynchronous announcements (unscheduled or unexpected news)*. Main stream news, rumours and social media normally arrive asynchronously in an unstructured textual form. A substantial portion of pre news arrive at pre scheduled times and generally in a structured form.

Scheduled (news) announcements often have a well defined numerical and textual content, and may be classified as structured data. These include macro economic announcements and earnings announcements. Macro economic news, particularly economic indicators from the major economies are widely used in automated trading. They have an impact in the largest and most liquid markets, such as foreign exchange, government debt and futures markets. Firms often execute large and rapid trading strategies. These news events are normally well documented, thus thorough backtesting of strategies is feasible. Since indicators are released to a precise schedule, market participants can be well prepared to deal with them. These strategies often lead to firms fighting to be first to the market; speed and accuracy are the major determinants of success. However the technology requirements to capitalise on events is substantial. Content publishers often specialise in a few data items and hence trading firms often multi source their data. Thomson Reuters, Dow Jones and Market News International are a few leading content service providers in this space.

Earnings are a key driving force behind stocks’ prices. Scheduled earnings announcement information are also widely anticipated and used within trading strategies. The pace of response to announcements has accelerated greatly in recent years. (See p.104-105 Leinweber) Wall Street Horizon and Media Sentiment (see Munz 2010) provide services in this space. These technologies allow traders to respond quickly and effectively to earnings announcements.

Event driven asynchronous news streams in unexpectedly over time. These news items usually arrive as textual, unstructured, qualitative data. It is characterised as being non-numeric and difficult to process quickly and quantitatively. Unlike analysis based on quantified market data textual news data contains information about the effect of an event and the possible causes of an event. However, to be applied in trading systems and quantitative models it needs to be converted to a quantitative input time series. This could be a simple binary series where the occurrence of a particular event or the publication of a news article about a particular topic is indicated by a one and the absence of the event can be indicated by a zero. Alternatively we can try to quantify other aspects of news, over time. For example, we could measure news flow (volume of news) or we could determine scores (measures) based on the language sentiment of text or determine scores (measures) based on the market’s response to particular language.

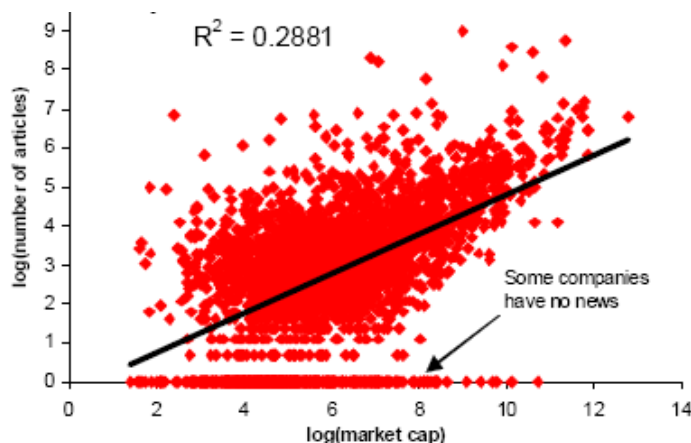
It is important to have access to historical data for effective model development and backtesting. Commercial news data providers normally provide large historical archives for this purpose. The details of historic news data for global equities provided by RavenPack and Thomson Reuters NewsScope are summarized in Appendix 1. They are taken from the RavenPack NewsScores User Guide (RavenPack 2010) and Thomson Reuters NewsScope Sentiment engine (2009).

2.2 Pre analysis of news data

Collecting, cleaning and analysing news data is challenging. Major news providers collect and translate headlines and text from a wide range of worldwide sources. For example, the Factiva database provided by

Dow Jones holds data from 400 sources ranging from electronic newswires, newspapers and magazines.

We note *there are differences in the availability of news data for different companies*. Larger companies (with more liquid stock) tend to have higher news coverage/news flow. Moniz, Brar and Davis (2009) sort companies by $\frac{\log \text{market cap}}{6 \text{ month ADV}}$. The top quintile 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 cap companies. See figure 2.



Source: RavenPack, Macquarie Capital (USA), May 2009

Figure 2: Number of news items versus log market capitalisation: Taken from Cahan, Jussa and Luo 2009

Classification of news items is important. Major newswire providers tag incoming news stories. A reporter entering a story, on to the news systems, will often manually tag it with relevant codes. Further 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. They reveal the stories' topic areas and other important meta data. For example, they may include information about which company a story is about. Tagged stories held by major newswire providers will also be accurately time stamped. The SEC is pushing to have companies file their reports using eXtensible Business Reporting Language (XBRL). Rich Site Summary (RSS) feeds (an XML format for web content) allow customised, automated analysis of news events from multiple online sources.

Tagged news stories provide us with hundreds of different types of events. We need to distinguish what types of news are relevant to our model (application). Further the market may react differently to different types of news. For example, Moniz et. al. (2009) finds the market seems to react more strongly to corporate earnings related news than corporate strategic news. They postulate that it is harder to quantify and incorporate strategic news into valuation models, hence it is harder for the market to react appropriately to such news.

Machine readable XML news feeds can turn news events into exploitable trading signals since they can be used relatively easily to backtest and execute event study based strategies. See Kothari and Warner (2005) and Campbell, Lo, and MacKinlay (1996) for in depth reviews of event study methodology. Leinweber (2010) uses Thomson Reuters tagged news data to investigate several news based event strategies. Elementised news feeds mean the variety of event data available is increasing significantly. News providers also provide archives of historic tagged news which can be used for backtesting and strategy validation. News event algorithmic trading is reported to be gaining acceptance in industry. (Schmerken 2006)

To apply news effectively in asset management and trading decisions *we need to be able to identify*

news which is both relevant and current. This is particularly true for intraday applications, where algorithms need to respond quickly to accurate information. We need to be able to identify an “information event”, that is, we need to be able to distinguish those stories which are reporting on old news (previously reported stories), from genuinely “new” news. As would be expected, Moniz et. al. (2009) finds markets react strongly when “new” news is released.

Tetlock, Saar-Tsechansky and Macskassy (2008) undertake an event study which illustrates the impact of news on cumulative abnormal returns (CAR). They use 350,000 news stories about S&P 500 companies appearing in the Wall Street Journal and Dow Jones News Service from 1984 - 2004. Each story’s (language) sentiment is determined using the General Inquirer and a story is classified as either positive or negative. The CAR for each story classification type, relative to the date of the news release is shown in Figure 3. There seems to be a connection between a news story’s release and the CAR. However, there also seems to be some “information leakage” since CAR seem to react before the date of the story’s release. Leinweber (2009) considers that this may be due to the inclusion of me-too stories that refer back to an original release of “new” news. This highlights that though textual news may have an obvious connection with returns it needs to be processed carefully and effectively.

Reuters identify relevance scores for different news articles. This measures by how much a measure of relevance the article is about a particular company. They also measure article novelty(uniqeness) which determines the repetition among articles and how many similar articles there are for a particular company. RavenPack (2010) also apply machine learning techniques to extract similar pertinent information for incoming newswire stories. In particular, they distinguish stories which are events. These are stories which carry the first mention of a particular theme. Stories which are not events are excluded. This is done to minimise the number of duplicate stories.

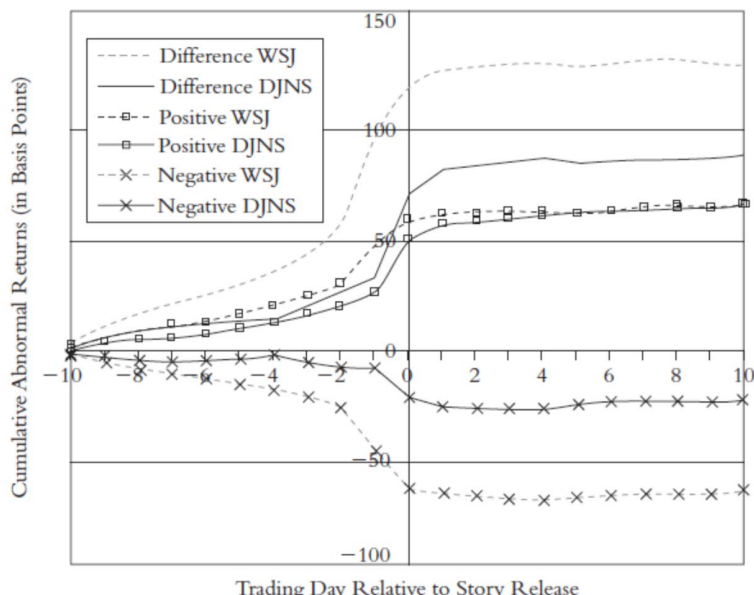


Figure 3: CAR start to respond several days before relevant news being published

Several studies also report *strong seasonality in newsflow* at hourly, daily and weekly levels. (Lo 2008 Hafez 2009, Moniz et. al. 2009) A valuable aspect of pre-analysis of news data is to identify periods of

unexpected newsflow levels, from periods of variation due to seasonality, in order to identify periods where significant levels of information are flowing into the market. Hafez (2009) investigates the seasonality patterns of news arrival. Figure 4 shows the intra day pattern. He notes that larger volumes of newsflow arrive just before the opening of the European, US and Asian trading sessions. On the intra week level we can see little newsflow takes place on the weekends. In the week, the peak of newsflow occurs on Wednesday and Thursday, while the trough falls on Friday. Lo also notes that the median number of weekday Reuters news alerts falls between 1,500 and 2,000, while the median for the entire weekend is 130.

The *time of the day when news is released*, has also been found to be relevant in understanding the connection between market variables and news. Robertson, Geva and Wolff (2006) find that there is a greater likelihood of events that lead to rising volatility at the start of the day. Boyd, Hu, and Jagannathan (2005) find that *market conditions* can influence the types of news that are reported. They report that interest rate information dominates in expansionary periods. In contrast information about future corporate dividends dominates when the markets are contracting.

As would be expected the *informational content of news* has a large influence on how markets react to news (Blasco, Corredor, Del Rio and Santamaria 2005, Boyd, Hu, and Jagannathan 2005 Liang 2005 and Tetlock 2007). We discuss how to extract the informational content of news (that is the sentiment) in section 3. It has been recognised that stock returns react more strongly to “negative” news than “positive” (Tetlock 2007). There also tends to be a positive sentiment bias, that is there is a larger volume of “positive” news to “negative” news. Das and Chen (2007) find that a histogram of normalised stock message board sentiment is positively skewed. There are days when messages about a stock are extremely optimistic but there is not a similar level of expression of pessimistic views. RavenPack (2010) also find a positive sentiment bias in their sentiment classifiers. This bias is more marked in bull markets than bear markets. They report a ratio of 2:1 of positive sentiment to negative sentiment stories in bull markets.

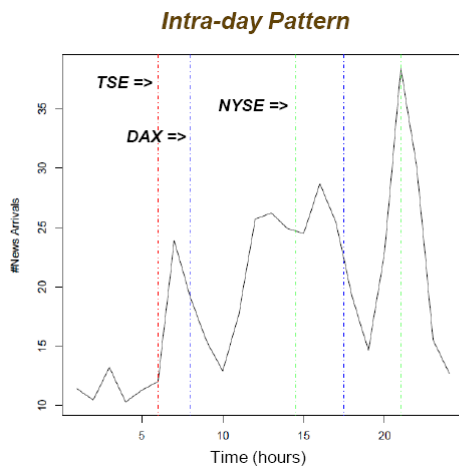


Figure 4:

The relationship of different news stories to each other is also an important consideration. Companies may make several announcements that fall under different classifications on the same day. These may or may not be related and may be related to varying degrees. For example a company may announce a profit warning, resignation of their CEO and provide guidance on its sales outlook. The dependence or independence between different news stories is a consideration.

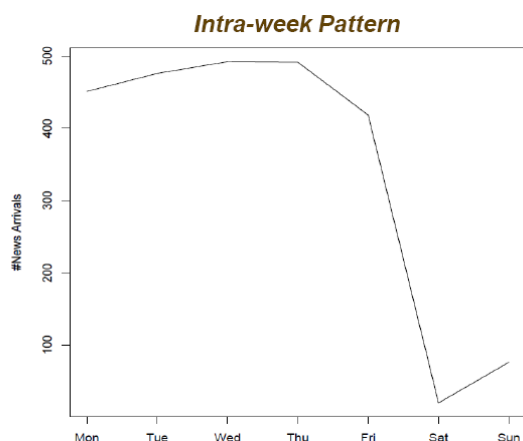


Figure 5:

3 Turning qualitative text into quantified metrics and time series

A salient aspect of news analysis is to discover the *informational content of news*. Converting qualitative text into a machine readable form is a challenging task. We may wish to distinguish whether a story’s informational content is positive or negative, that is, determine its sentiment. We may go further and try to identify “by how much” the story is positive or negative. In doing this we may try to assign a quantified sentiment score or index to each story. A major difficulty in this process is identifying the context in which a story’s language is to be judged. Sentiment may be defined in terms of how positively or negatively a human (or group of humans) interprets a story. That is the emotive content of the story for that human. In particular standards can be defined using experts to classify stories. Some of RavenPack’s classifiers are calibrated using finance experts to define the context of language. Further dictionary based algorithms which use psychology based interpretations of words may be used. Since different groups of people are effected by events differently and have different interpretations of the same events, conflicts may arise. Moniz et. al. (2009) gives an example of the term “dividend cuts”. This may be classified as a negative term by a dictionary based algorithm. In contrast, it may be interpreted positively by market analysts who may believe this indicates the company is saving money and more able to repay its debts. Loughran and McDonald (2010) also consider how context effects how the tone of text is interpreted. They note a Psychological dictionary like the Harvard-IV-4 may classify words as negative when they do not have a negative financial meaning. They develop an alternative negative word list that better reflects the tone of financial text.

An attractive alternative is to use market based measures to interpret and define the importance of news. The markets’ relative change in returns or volatility, for a particular asset or asset class lagged against a relevant news story, can be used to define the sentiment (informational content) of the news story. This approach intrinsically assumes that the market has responded to the news story. Lo (2008) uses this approach for creating the Reuters Newscope Event Indices. He creates separate indices for market responses to news, in terms of returns and volatility. So he assumes that sentiment measured in the context of these two variables are different. This approach is quite pragmatic and is focussed on using the news content directly in the context that the modeller is interested in. Lavernko, Schmill, Lawrie, Ogilvie, Jensen and Allan (2000), Moniz et. al. (2009), Peramunetilleke and Wong (2002) and Luss and d’Aspremont (2009) also use market based measures in determining the “sentiment” of news. SemLab (see Vreijling 2010) provide a

tool which allows the user to filter news items and examine each item's impact on market variables. Using this interactive tool, the user is able to define their own tailored context of "sentiment".

Given a definition of sentiment, machine learning and natural language techniques are frequently used to determine the sentiment of new incoming stories. Hence we can determine sentiment scores over time as news arrives. Such sentiment indices then allow us to develop systematic investment and risk management processes. Once a sentiment index is constructed, to use it effectively, we must be able to find evidence of a relationship with relevant asset returns, trading volumes or volatility.

The definition of market sentiment is very much context dependent. In general we are interested in discovering the "informational context of news" In this review paper for the purpose of (quantitative) modelling applications we use these two terms "news sentiment" and "informational content of news" interchangeably, and in this section we discuss some of the leading methods of computing / quantifying "sentiment" and other related measures.

We review below Das and Chen (2007) and Lo (2008). Both paper cover the following items.

1. A definition of the context of sentiment is given.
2. Application of algorithms (natural language, machine learning and linear regression) to calibrate and define sentiment scores.
3. Validation of the effectiveness of the scores by comparing their relationship with relevant asset returns, volumes or volatility.

Das and Chen (2007) use statistical and natural language techniques to extract investor sentiment from stock message boards and generate sentiment indices. They apply their method for 24 technology stocks present in the Morgan Stanley High Tech (MSH) Index. A web scraper program is used to download tech sector message board messages. Five algorithms, each with different conceptual underpinnings are used to classify each message. A voting scheme is then applied to all five classifiers.

Three supplementary databases are used in the classification algorithms.

1. "*Dictionary*" is used for determining the nature of the word. For example, is it a noun, adjective or adverb?
2. "*Lexicon*" is a collection of hand picked finance words which form the variables for statistical inference within the algorithms.
3. "*Grammar*" is the training corpus of base messages used in determining the in-sample statistical information. This information is then applied for use on the out-of-sample messages.

The lexicon and grammar jointly determine the context of the sentiment. Each of the classifiers relies on a different approach to message interpretation. They are all analytic, hence computationally efficient.

1. *Naive classifier* (NC) is based on a word count of positive and negative connotation words. Each word in the lexicon is identified as being positive, negative or neutral. A parsing algorithm negates words if the context requires it. The net word count of all lexicon matched words is taken. If this value is greater than one, we sign the message as a buy. If the value is less than one the message is a sell. All others are neutral.
2. *Vector distance classifier* Each of the D words in the lexicon is assigned a dimension in vector space. The full lexicon then represents a D-dimensional unit hypercube and every message can be described as a word vector in this space ($m \in \mathbb{R}^D$). Each hand tagged message in the training corpus (grammar) is converted into a vector G_j (grammar rule). Each (training) message is pre classified as positive,

negative or neutral. We note that Das and Chen use the terms Buy/Positive, Sell/Negative and Neutral/Null interchangeably. Each new message is classified by comparison to the cluster of pretrained vectors (grammar rules) and is assigned the same classification as that vector with which it has the smallest angle. This angle gives a measure of closeness.

3. *Discriminant based classification* NC weights all words within the lexicon equally. The discriminant based classification method replaces this simple word count with a weighted word count. The weights are based on a simple discriminant function (Fisher Discriminant Statistic). This function is constructed to determine how well a particular lexicon word discriminates between the different message categories ($\{ \text{Buy, Sell, Null} \}$). The function is determined using the pre classified messages within the grammar. Each word in a message is assigned a signed value, based on its sign in the lexicon multiplied by the discriminant value. Then as for NC a net word count is taken. If this value is greater than 0.01, we sign the message as a buy. If the value is less than -0.01 the message is a sell. All others are neutral.
4. *Adjective - adverb phrase classifier* is based on the assumption that phrases which use adjectives and adverbs emphasize sentiment and require greater weight. This classifier also uses a word count but uses only those words within phrases containing adjectives and adverbs. A “tagger” extracts noun phrases with adjectives and adverbs. A lexicon is used to determine whether these significant phrases indicate positive or negative sentiment. The net count is again considered to determine whether the message has negative or positive overall sentiment.
5. *Bayesian classifier* is a multi variate application of Bayes Theorem. It uses the probability a particular word falls within a certain classification and is hence indifferent to the structure of language. We consider three categories $C = 3 \quad c_i \quad i = 1, \dots, C$. Denote each message $m_j \quad j = 1, \dots, M$. The set of lexical words is $F = \{w_k\}_{k=1}^D$. (The total number of lexical words is D) We can determine a count of the number of times each lexical item appears in each message $n(m_j, w_k)$. Given the class of each message in the training set we can determine the frequency with which a lexical word appears in a particular class. We are then able to compute the conditional probability of an incoming message j falling in category i , $Pr(m_j|c_i)$, from the word based frequencies. $Pr(c_i)$ is set to the proportion of messages in the training set classified in class c_i . For a new message we are able to compute the probability it falls within class c_i given its component lexicon words, that is $P(c_i|m_j)$, through an application of Bayes Theorem. The message is classified as being from the category with the highest probability.

A voting scheme is then applied to all five classifiers. The final classification is based on achieving a majority amongst the five classifiers. If there is no majority the message is not classified. This reduces the number of messages classified but enhances the classification accuracy.

Das and Chen also introduce a method to detect message ambiguity. Messages posted on stock message boards are often highly ambiguous. The grammar is often poor and many of the words do not appear in standard dictionaries. They note “Ambiguity is related to the absence of “aboutness””. The General Inquirer has been developed by Harvard University for content analyses of textual data. They use it to determine an independent optimism score for each message. By using a different definition of sentiment it is ensured there is no bias to a particular algorithm. The optimism score is the difference between the number of optimistic and pessimistic words as a percentage of the total words in the body of the text. This score allows us to rank the relative sentiment of all stories within a classification group. For example, they can rank the relative optimism of all stories which have been classified by their scheme as positive. The mean and standard deviation of the optimism score for different classification types ($\{ \text{Buy, Sell, Null} \}$) can be calculated. They filter *in* and consider only highly optimistically scored stories in the positive category. For example only those stories with optimism scores above the mean value plus one standard deviation are considered. Similarly they filter *in* and consider only the most highly pessimistic scores in the negative category. Once the classified stories are further filtered for ambiguity, it is found that the number of false positives dramatically declines.

Once the sentiment for each message is determined using the voting algorithm, a daily sentiment index is compiled. The classified messages up to 4pm each day are used to create the aggregate daily sentiment for each stock. A buy (sell) message increments (decrements) the index by one. These indices are further aggregated across all stocks to obtain an aggregate sentiment for the technology portfolio. A disagreement measure is also constructed

$$DISAG = \left| 1 - \left| \frac{B - S}{B + S} \right| \right| \quad (1)$$

B (S) is the number of buy (sell) messages. This measure lies between 0 (no disagreement) and 1 (high disagreement) and is computed as a daily time series. The daily MSH index and component stock values are also collected. Trading volatility and volume of stocks are also calculated and message volume is also recorded. All the time series are normalised.

Das and Chen check that the constructed sentiment indices have a relationship with relevant asset variables. The relationship between the MSH index and the aggregate sentiment index is investigated. Fig 2 plots the two against each other. These two series do seem to track each other. The sentiment index is found to be highly autocorrelated out to two trading weeks. Regression analysis is undertaken to investigate the relationship. They conclude sentiment does offer some explanatory power for the level of the index. However, the autocorrelation makes it difficult to establish the empirical nature of the relationship.

Das and Chen undertake regression analysis between the individual stock level and the individual stock sentiment level and find there is a significant relationship. (t-statistic of coefficient falls within a significant level) The relationship between first differences is much weaker. We cannot conclude there is a strong predictive ability on forecasting individual stock returns. Sentiment and stock levels are not unrelated, but determining the precise nature of the relationship is difficult.

Fig 4 summaries the relationship between the sentiment measure, disagreement measure, message volume, trading volume and volatility. Sentiment is inversely related to disagreement. As disagreement increases the sentiment falls. Sentiment is correlated to high posting volume. As discussion increases this indicates optimism about that stock is rising. There is a strong relationship between message volume and volatility. This is consistent with Antweiler and Frank (2002). Trading volume and volatility are strongly related to each other.

Lo (2008) develops the Reuters NewsScope Event Indices (NEI) which are constructed to have “predictive” power for particular asset returns and (realised) volatility. They are constructed in an integrated framework where news, returns and volatility are used in calibrating the indices. The white paper (dated November 2007) considers specifically indices for foreign exchange. However, the method can be applied to other asset classes.

Lo uses news alerts in developing his sentiment indices. These are quick news flashes which are issued when a newsworthy event occurs. They are both timely and relevant. An example of Reuters NewsScope Alert

TimeStamp 02 AUG 2007 04:44:26.155

Alert Tsunami Warning Issued for Japan’s Western Hokkaido Coast

Tags JP ASIA NEWS DIS LEN RTRS

The alerts comprises three items (i) TimeStamp (ii) A short headline and (iii) Tags and meta data. The tags are machine readable and will often contain information about the topic area. The headlines lend themselves well to machine analysis since they are concise and formed from a small vocabulary. Lo notes the purpose of the indices is to rapidly identify and report market moving information. Once constructed he undertakes (event study) experiments to validate their quality, developing metrics which have the potential to indicate whether the indices are able to predict significant market movements.

Framework for real time news analytics

We consider here the framework for developing the Reuters NEI. For a given asset class and related topic area the following parameters are used.

- (1) List of keywords and phrases with real valued weights; $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$.
- (2) A rolling “sentiment” window of size r (say 5/10 minutes).
- (3) A rolling calibration window of size R (say 90 days).

Initially a *raw score* is created.

We have $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$, where W_1 is the first keyword and γ_1 is the weighting for the first keyword.

The raw score at time t is assigned by considering the time period $(t - r, t]$. (w_1, \dots, w_k) is the vector of keyword frequencies in $(t - r, t]$, that is, w_i is the number of times keyword W_i occurred in the last r minutes. The raw score is defined as

$$s_t \equiv \sum_i \gamma_i w_i \quad (2)$$

The raw score will tend to be high when the news volume is high. A *normalised score* is therefore produced using the rolling calibration window. At all times t for the R days in the calibration window, we record

- (i) the raw score s_t that would have been assigned,
- (ii) the news volume; $n_{[t-r,t]}$ the number of words that were observed in the time interval $[t - r, t)$.

The normalised score is determined by comparing the current raw score against the distribution of raw scores in the calibration window, where the news volume equalled the current news volume. This means we only consider those raw scores where the news volume equals the current news volume.

$$S_t \equiv \frac{|\{t' \in [t - R, t) : n_{[t'-r,t')} = n_{[t-r,t)} \& s_{t'} < s_t\}|}{|\{t' \in [t - R, t) : n_{[t'-r,t')} = n_{[t-r,t)}\}|} \quad (3)$$

We notice the numerator is a subset of the denominator, hence $S_t \leq 1$. If $S_t = 0.92$, we can say 92% of the time when news volume is at the current level, the raw score is less than it currently is. Lo creates an alternative score based on topic codes. Instead of counting word frequencies, the fraction of news alerts (in the last r minutes) tagged with particular topic codes, are used.

Naturally the scoring method is dependent on the list of keywords/topic areas (W_1, \dots, W_k) and the real valued weights $(\gamma_1, \dots, \gamma_k)$. The lists of keywords/topics were created by selecting the major news categories that related to the asset class (foreign exchange) and creating lists, by hand, of words and topic areas that suggest news relevant to the categories. A tool was created to extract news from periods where high scores were assigned. This news was then manually inspected, so that the developer could determine whether the keywords (topics) were legitimate or needed adjusting.

The optimal weights $(\gamma_1, \dots, \gamma_k)$ for the intraday return sentiment index were determined by regressing the word (topic) frequencies against the intraday asset returns. Similarly the (optimal) weights for the intraday volatility sentiment index were determined by regressing the word (topic) frequencies against the intraday (de-seasonalised) realised volatility. Volatility was observed to show strong seasonality on intraday timescales, hence this series was de-seasonalised prior to derivation of the weights. Returns did not exhibit any seasonality. The time series are given on an intraday basis, hence to keep the data manageable a random subset of the observations are used in calibration. Lo notes the determination of the weights can be expressed as a more general classification problem. Other techniques might be applied, in particular machine

learning algorithms such as the perceptron algorithm or support vector machines. He suggests further study is required to find the best approach, but the standard linear regression approach does perform well.

To establish that the final NEI have empirical significance, Lo undertakes detailed event study analysis. He uses the NEI series to define an event. An event is defined to take place when the index exceeds a certain threshold (say 0.995). He then removes any events that follow in less than one hour of another event. This guards against identifying “new” events which are actually based on old news. The behaviour of exchange rates before and after these events are then studied. Two time series are considered; the log returns and the deseasonalised squared log returns. He then tests the null hypothesis that the distribution of log returns / deseasonalised squared log returns are the same before and after the events. He uses samples of one hour centered on the events.

We can visually assess the impact of events on the volatility of EUR/USD exchange rate.

(1) Figure 6 shows the averaged volatility event window. The pre event (averaged) volatility is shown in blue, and the post event (averaged) volatility is shown in red. There is a peak at the centre where there is a significant increase in volatility.

(2) Figure 7 shows the density function of pre event samples and post event samples of deseasonalised squared log returns. The shift to the right indicates an upward shift in volatility on average.

As well as visual inspection, statistical tests can be introduced to compare the pre and post event samples. A t-test can be used to test equality of the means in the two samples. Levene’s test can be used to determine whether there has been a change in the standard deviation. The χ^2 goodness of fit test can be used to determine whether the two samples are likely to have come from different distributions.

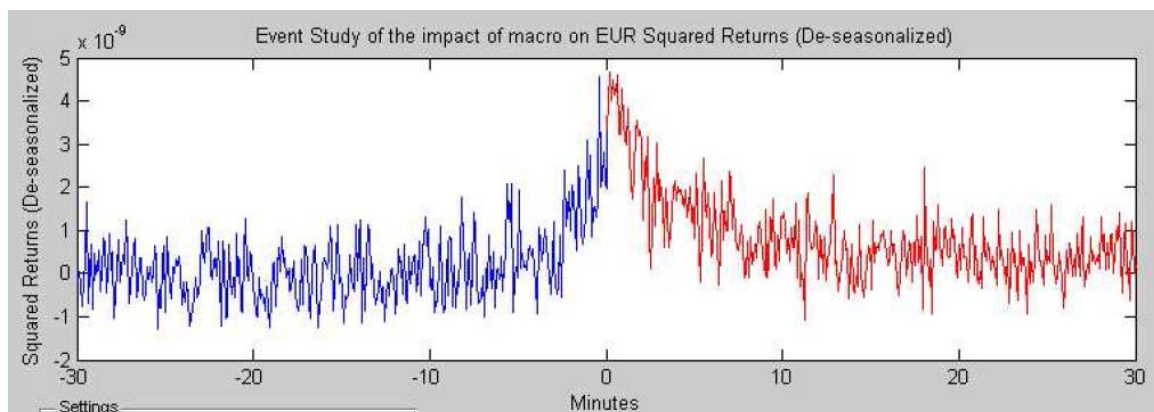


Figure 6: Pre and post event squared returns

Indices and FX Implied Volatility

Lo finds that the event studies confirm the constructed event indices, on average, impact the *realised* foreign exchange volatility. He further considers the relationship of the indices to *implied* volatility. The NEI volatility indices are constructed to predict volatility over 30 minute periods. Implied volatility gives the markets’ expectations of volatility over a much longer horizon, typically 30 days. Event study analysis

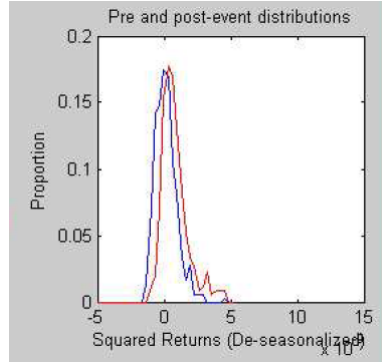


Figure 7: Distribution of pre and post event squared returns

between implied volatility and the NEI volatility indices shows no evidence of a relationship. Lo feels that implied volatility and the indices may function as complementary sources of information for risk management, since they intrinsically focus on different time horizons.

4 Models and applications

News Analytics in finance is the use of technology and algorithms to process news, within the investment management process. It allows investors to update their beliefs about the future market environment more effectively. This technology may be geared towards human decision support or it may be used to create automated quantitative strategies. The use of news data in addition to historic market data makes models more proactive and less reactive. The applications broadly fall into two areas; trading and risk control.

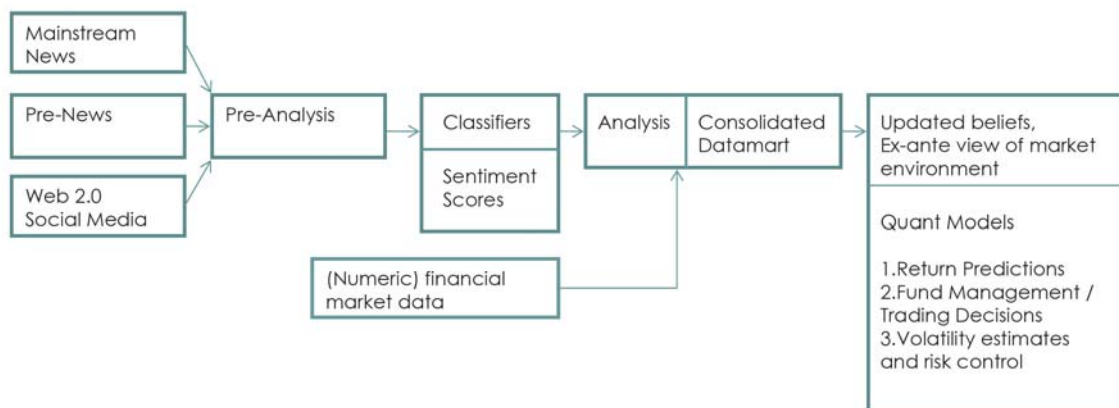


Figure 8: Information flow and computational architecture

4.1 Information flow and computational architecture

News analytics in finance focus on improving IT based legacy system applications. These improvements come through research and development directed to automating/semi automating programmed trading, fund rebalancing and risk control applications.

The established good practice of applying these analytics in the traditional manual approach are as follows. News stories and announcements arrive synchronously and asynchronously. In the market assets' (stocks, commodities, FX rates, etc) prices move (market reactions). The professionals digest these items of information and accordingly make trading decisions, investment decisions, recompute their risk exposures.

The information flow and the (semi) automation of the corresponding IS architecture is set out in figure 8. There are two streams of information which flow simultaneously, news data and market data. Pre-analysis is applied to news data; it is further filtered and processed by classifiers to relevant metrics. This is consolidated with the market data of prices together they constitute the classical datamart which feed into whatever relevant model based applications are sought. A key aspect of these applications is that they set out to provide technology enabled support to professional decision makers thereby achieve intelligence amplification (Leinweber 2009).

4.2 Trading and fund management

Generally traders and quantitative fund managers seek to identify and exploit asset mispricings, before they correct, in order to generate alpha. Most simply they may use (quantified) news data to *rank stocks* and identify which stocks are relatively attractive (unattractive). They may then buy (sell) the highest (lowest) ranking stocks, thereby rebalancing a portfolio composed of desired weights on the selected stocks. Similarly the news data may be used to identify trading signals for particular stocks. Alternatively analysts may use *factor models* to process new sources of news data. (Factor models, which are applied to give updated estimates of future asset returns and volatility, allow us to determine an optimal future portfolio to hold. That is, they tell us which assets to hold and also in what proportions.) Analysts may also use news data to identify and exploit *behavioural biases* in investor behaviour arising due to the market and analysts' misreaction to new information. In particular this can arise due to delayed information diffusion or due to investors' inattention and limited ability to process all relevant information instantaneously.

Stock picking and ranking

Li (2006) uses a simple ranking procedure to identify stocks with positive and negative (financial language) sentiment. He examines form 10-K Securities and Exchange Commission (SEC) filings for non-financial firms between 1994-2005. He creates a "risk sentiment measure" which is formed by counting the number of times the words risk, risks, risky, uncertain, uncertainty and uncertainties occur within the management discussion and analysis sections. A strategy which goes long in stocks with a low risk sentiment measure and short stocks with a high risk sentiment measure is found to produce a reasonable level of returns. Leinweber (2009) notes it is rumoured similar approaches are being applied. The performance of the strategy deteriorates in recent years, possibly due to wider use of such strategies.

Moniz et. al. (2009) focuses on turning news signals into a trading strategy. Equity analysts collect, process and disseminate information on companies to investors. In particular they use their research to form earnings forecasts for companies. Earnings momentum strategies, thus proxy for corporate newsflow. Moniz notes these strategies do not explicitly identify the piece of information that has triggered the change in earnings forecast. He investigates whether news leads earnings revisions. He finds news data can be used to reinforce proxies for news already incorporated in models and a strategy based on earnings momentum reinforced by newsflow is found to be effective.

Event studies based on news events can also allow fund managers to identify potentially under/over priced stocks. See the discussion in section 2.2.

Factor models

The Efficient Market Hypothesis (EMH) asserts that financial markets are “informationally efficient” so prices of traded assets reflect all known information and update instantaneously to reflect new information. Further it is assumed that agents act rationally. It is widely accepted, within the fund management and trading community, that the EMH, particularly in its strong form, does not hold. In the long term markets may be efficient. But “The long run is a misleading guide to current affairs. In the long run we are all dead.” as John Maynard Keynes said. In the shorter term traders and quantitative fund managers seek to identify and exploit asset mispricings, before these prices correct themselves, in order to generate alpha. In undertaking this process they often seek to gain a competitive advantage by applying improved and differentiating sources of data and information.

The capital asset pricing model (CAPM) is the classical approach to pricing equities (Sharpe 1964 and Lintner 1965). Any asset’s return can be split into a component that is correlated with the market’s return and a residual component that is uncorrelated with the market. Under the CAPM, it is assumed that, the expected return for the residual component is zero and any stock’s expected return is dependent only on the expected return of the market. The CAPM states that only risk (uncertainty) due to market variability should be priced. Residual risk can be diversified and therefore should not be compensated.

The arbitrage pricing theory (APT) (introduced by Ross (1976)) extends the CAPM to a more general linear model where additional sources of information to market returns are considered. Under the APT (multifactor models) an asset’s expected return is represented as a linear sum of several “risk” (uncertainty) factors that are common to all assets and an asset specific component. The APT states the investor should be compensated for their exposure to all sources of (non-diversifiable) risk.

Active portfolio managers seek to incorporate their investment insight to “beat the market”. An accurate description of asset price uncertainty is key to the ability to outperform the market. Tetlock, Saar-Tsechansky and Macskassy (2008) note that an investor’s perceptions about the future asset returns are determined by their knowledge about the company and its prospects, that is, by their “information sets”. They note that these are determined from three main sources: analysts forecasts, quantifiable publicly disclosed accounting variables and linguistic descriptions of the firm’s current and future profit generating activities. If the first two sources of information are incomplete or biased, the third may give us relevant information for equity prices.

Multi factor models are now widely used by fund managers in constructing alpha generating strategies (Rosenberg, Reid and Lanstein 1985). Identifying the relevant factors (and betas) is a measure of skill. Fund managers are always seeking new sources of advantage. This can be data and factors which translate to “quantitative knowledge”. “Profits may be viewed as the economic rents which accrue to [the] competitive advantage of ... superior information, superior technology, financial innovation” (Lo 1997). A “quantcentration” effect is frequently observed. That is, since most fund managers have access to the same sources of data, it is difficult to distinguish between their models and performance. Cahan, Jussa and Luo (2009) find that news sentiment scores provided by RavenPack act as an orthogonal factor to traditional quantitative factors currently used. Hence they add a diversification benefit to traditional factor models. In particular they note the value of this source of information during the credit crisis, when determining fundamentals (which traditional quant factors are based on) was problematic.

Behavioural biases

Behavioural economists challenge the assumption that market agents act rationally. Instead they propose that individuals display certain biased behaviour, such as, loss aversion (Kahneman and Tversky 1979), overconfidence (Barber and Odean 2001), overreaction (DeBondt and Thaler 1986) and mental accounting (Tversky and Kahneman 1981). Due to individual behavioural biases investors systematically deviate from optimal trading behaviour (Daniel, Hirshleifer and Teoh 2002, Hirshleifer 2001, Odean and Barber 1998). Behavioural economists use these biases to explain abnormal returns, rather than risk based explanations.

Naturally investor behaviour is dependent on individual and group psychology. Some of the research within behavioural finance seeks to understand the mechanisms of human investor behaviour, drawing heavily on the fields of neuroscience and psychology (See for example Peterson 2007). Lo (2004) proposes a new framework the Adaptive Market Hypothesis (AMH) which seeks to reconcile market efficiency with behavioural alternatives. This is an evolutionary model, where individuals adapt to a changing environment via simple heuristics.

As noted before the relationship between news and markets is complex. A number of studies consider how investors react to news releases, in particular, the behavioural and cognitive biases in their reactions to news. Quantitative investors often seek to systematically exploit the anomalies observed in prices arising from investors' behavioural biases (Moniz et. al. 2009 Barber and Odean 2007 Seasholes and Wu 2004). There is a commercial fund called MarketPsy which employs strategies that exploit "collective investor misbehaviour" (see <http://marketpsy.com/>).

Barber and Odean (2007) consider evidence for the behavioural bias that individual investors have a tendency to buy attention-grabbing stocks. Attention grabbing stocks are defined as ones that display abnormal trading volumes, extreme one day returns or are mentioned on the Dow Jones New Service. In contrast professional managers who are better equipped to assess a wider range of stocks are less prone to buying attention grabbing stock. In particular institutional investors, who use computers to manage their searches, normally specialise in a particular sector and may consider only those stocks that meet certain criteria. For every buyer there must be a seller. So if one group incurs losses the other group profits. If individual investors fail to react appropriately to news and attention, there is scope for institutional investors to profit. Seasholes and Wu (2004) find individual investors tend to buy stocks that hit an upper price limit. They find an impact on the prices of these attention grabbing stocks, which reverses to pre event levels within ten working days. Further they find a group of professional investors who profit from the biased behaviour of the individual investors.

Fang and Peress (2009) consider whether media coverage can help predict the cross-section of future stock returns. They find stocks with no media coverage outperform widely covered stocks even after allowing for well-known risk factors. This is contrary to the findings of Barber and Odean (2007). But this finding supports Merton's (1987) investor recognition hypothesis. Da, Engleberg and Gao (2009) also consider how the amount of attention a stock receives affects its cross-section of returns. They use the frequency of Google searches for a particular company as a measure of the amount of attention a stock receives. They find some evidence that changes in investor attention can predict the cross-section of returns. This is most pronounced amongst the small cap stocks.

Some researchers consider how informational flows cause investors to update their expectations, in order to explain momentum and reversal effects. DeBondt and Thaler (1986) suggest that investors overreact to recent earnings placing less emphasis on long term averages. Daniel, Hirshleifer, and Subrahmanyam (1998) suggest price momentum is a result of investors overreacting to private information causing prices to be pushed away from fundamentals. In contrast Hong and Stein (1999) suggest price momentum occurs due to investors underreacting to new information. They suggest information diffuses slowly and is gradually incorporated into prices. Hirshleifer, Lim and Teoh (2010) find that when there is a significant number of earnings announcements in the market, investors are distracted and underact to relevant new information and the post announcement drift is strong. Investors fail to price the information efficiently, leaving an opportunity for quantitative investors. Scott, Xu and Stumpp (2003) conclude that price momentum is caused by under reaction of stocks to earnings related news. This is contrary to prior literature which suggested that price momentum was connected to trading volume.

Chan (2003) finds stocks with major public news exhibit momentum over the following month. In contrast stocks with large price movements, but an absence of news, tend to show return reversals in the following month. This would support a trading strategy based on momentum reinforced with news signals. Da, Engleberg and Gao (2009) extend their analysis of Google searches to consider the debate on how momentum works. They find price momentum is stronger in stocks with high levels of Google (SVI) searches. This

supports Daniels et al. (1998) view since one would expect investors to overreact to stocks they are paying close attention to. Gutierrez and Kelley (2007) Hou, Peng and Xiong (2009) also investigate the relationship between news(information flows) and momentum.

4.3 Monitoring risk and risk control

For effective financial risk control companies need to identify, understand and quantify potential (adverse) outcomes, their related probabilities and the severity of their impacts. This knowledge allows them to assess how best to manage and mitigate risk. Traditionally historic asset price data has been used to estimate risk measures. These traditional approaches have the disadvantage that they provide ex post retrospective measures of risk. They fail to account for developments in the market environment, investor sentiment and knowledge. Incorporating measures or observations of the market environment within the estimation of future portfolio return distributions is important, since the market conditions are likely to vary from historic observations. This is particularly important when there are significant changes in the market. In these cases risk measures, calibrated using historic data alone, fail to capture the true level of risk (See Mitra, Mitra and diBartolomeo 2009 and diBartolomeo and Warrick 2005). Recent technological developments have enabled the creation of data-mining tools that can interpret live news feeds. (See section 3 also RavenPack 2010; Brown/Thomson Reuters 2010; Vreijling/SemLab 2010) Mitra, Mitra and diBartolomeo 2009 find that updating risk estimates using news data can provide dynamic (adaptive) measures that account for the market environment. Further these measures may be useful in identifying and giving early ex-ante warning of extreme risk events.

The risk structure of assets may change over time, in response to news. Patton and Verardo (2009) investigate whether the systematic risk (beta) of stocks increases in response to firm specific news (in the form of earnings announcements). They undertake an event study on the beta of stocks around their earnings announcement dates. The change in beta on announcement date is decomposed into the change due to an increase in volatility of that stock and the change due to an increase in the covariance with the index. They find that news releases do have an important impact on the risk of stocks. Further much of the beta increase arises from an increase in covariance with other stocks. This suggests there could be a contagion effect in the information releases for one stock on the price movements of other stocks. This supports anecdotal evidence that investors will monitor earnings of related stocks when investigating the earnings of a particular stock. They suggest the credit crisis (2008) could be viewed as a negative earnings surprise for the market. Correlations were observed to increase during this period.

The relationship between public information release and asset price volatility has been widely investigated and noted. Ederington and Lee (1993) find a relationship between macroeconomic announcements and foreign exchange and interest rate futures return volatility. Graham, Nikkinen and Sahlstrom (2003) find stock prices on S&P500 are also influenced by macro economic announcements. Kalev, Liu, Pham, and Jarnecic (2004) find that a GARCH model for equity returns which incorporates asset specific news gives improved volatility forecasts. This study is extended in Kalev and Duong (2010). Robertson, Geva and Wolff (2007) also consider a GARCH model which accounts for “content aware” measures of news.

It is observed that volatility is higher in down markets. This is sometimes referred to as the “*leverage effect*”. Dzielinski et al. (2010) refer to it as *volatility asymmetry*. Their investigation concludes it is likely to be driven by the over reaction of private investors to bad news. In line with this theory, they find an increase in private investor attention to negative news can predict a rise in volatility. Increased private investor attention to negative news, is measured by a change in the level of Google searches for negative words related to the macro economy, such as recession.

The relationship between equity price volatility and web activity has also been widely investigated. Wysocki (1999) finds that spikes in Yahoo! Message board activity are good predictors of equity volatility (also volume and excess returns). Antweiler and Frank (2004) also have similar findings for equity volatility. An application for traffic analysis from the web was developed by Codexa for Bear Wagner to aid their risk

management strategy in predicting (unexpected) high volatility (Leinweber (2009) Ch10 p.237).

As discussed in section 3 Lo (2008) creates event indices (scores) that are constructed to predict changes in (foreign exchange) volatility. Empirical event studies show these are effective at converting incoming qualitative text (textual news) into quantitative signals that do indicate changes in volatility.

News data (flows) can also be used for non-quantitative risk control. Wolf detectors (circuit breakers) are a risk control feature for algorithmic trading built on machine readable news. Essentially they “break the circuit” stopping an automated algorithm from trading on a certain asset when particular types of news are released. It is important to try not to shout “Wolf!” when no wolf has actually appeared. These risk control features can be customized to only be tripped when substantive news events have occurred. Alternatively the algorithms can be turned back after the nature of the news has been programmatically analysed. This can be done using different features of machine readable news data (A Team 2010).

4.4 Desirable industry applications

Stock picking, trading and fund management(section 4.2) and risk control (section 4.3) are the established application areas in finance industry and the use of NA is researched to achieve improved performance.

- Market surveillance

Responding to the state of the market and taking into consideration the preoccupation of the watch-dogs, that is, the regulators market surveillance is becoming an important application area of quant models. It is gaining in importance because managers through internal control functions as much as external compliance requirement wish to have surveillance in place to catch rogue trading, insider information based trading. An innovative application of NA is to spot patterns which captures these.

- Trader decision support

News data can aid traders in making decisions. News data signals may confirm traders existing analysis or it may cause them to reconsider their analysis.

- Wolf detection / circuit breaker

Wolf detectors (circuit breakers) are a risk control feature for algorithmic trading built on machine readable news. Essentially they “break the circuit” stopping an automated algorithm from trading on a certain asset when particular types of news are released. It is important to try not to shout “Wolf!” when no wolf has actually appeared. These risk control features can be customised to only be tripped when substantive news events have occurred. Alternatively the algorithms can be turned back after the nature of the news has been programmatically analysed. This can be done using different features of machine readable news data. (See A-Team report)

- News flow algorithms

It is widely recognized that newsflow is a good indicator of volume and volatility. As the flow of news about a company rises, the volume traded rises resulting in more stock price volatility. If news flow can be used effectively to predict volume or volatility spikes then algorithms based on News VWAP versus VWAP may add value for trade execution strategies.

- Post trade analysis

Assist in proving best execution and trader performance??

News data is likely to add value for investors trading at all frequencies from volatility based strategies to equity trading.

- Alpha generating signal

News data can be used in alpha generation at various trading frequencies. News sentiment data may be used within factor models. Cahan, Jussa and Luo (2009) consider such an application. Their results

are positive and they find that such an approach does add value. In particular they note the value of this source of information during the credit crisis, when determining fundamentals (which traditional quant factors are based on) was problematic. News data can also aid quant investors to identify non rational biased behaviour of investors. These can then be exploited.

[? Tetlock, Saar-Tsechansky, and Macskassy (2008) note that an investor's perceptions about the future asset returns are determined by their knowledge about the company and its prospects, that is, by their "information sets". They note that these are determined from three main sources: analysts forecasts, quantifiable publicly disclosed accounting variables and linguistic descriptions of the firm's current and future profit generating activities. If the first two sources of information are incomplete or biased, the third may give us relevant information. ?]

- Stock screening tool

News data can be used to aid stock screening. In particular sentiment data may be used to try to guess the directional movement of future returns. Very good news stocks (for example top sentiment quintile) might be selected to be held long and very bad news stocks (for example bottom sentiment quintile) might be selected to be held short.

- Fundamental research

News analysis tools may aid traditional non-quant managers, by allowing them to undertake market research more efficiently.

- Risk Management

The use of news data within risk forecasting can allow for dynamic (adaptive) risk management strategies that are forward looking and are based on changing market environments. Further this risk analysis applied using news data can help investors understand event risk and how different kinds of events can impact their portfolio risk profile.

We may use certain news data within quantitative models. We may use it simply to forecast the directional impact of news on asset prices. In more sophisticated models we might wish to determine return predictions. Models which forecast volatility and volume on the basis of news will also find important applications within the investment management process. Volatility prediction for volatility traders?? Volume prediction for factor models that use volume as a factor?

5 Summary and discussions

The development of news analytics and its applications to finance through sentiment analysis is gaining progressive acceptance within the investment community. A growing number of academic studies have been conducted; in this paper we have reviewed these in a summary form. Research by service providers of data and content for the finance industry is also discussed in this paper and we have identified the applications of News Analytics to high frequency and low frequency trading as well as in risk control and compliance. The study of News Analytics draws upon research from a number of disciplines including natural language processing, AI pattern recognition and classifiers, text mining, information engineering as well as financial engineering; we believe News Analytics will soon become an important area of study within financial analytics.

6 References

References

- [1] W. Antweiler and M. Frank. Is All That Talk Just Noise? The Information Content of Stock Message Boards. *Journal of Finance*, 59(3), 2004.
- [2] L.S. Bamber, O.E. Barron, and T.L. Stober. Trading volume and different aspects of disagreement coincident with earnings announcements. *The Accounting Review*, 72:575–597.
- [3] B.M. Barber and T. Odean. Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment. *Quarterly Journal of Economics*, 116(1):261–292, 2001.
- [4] B.M. Barber and T. Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21(2):785–818, 2008.
- [5] B.M. Barber and T. Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors UPDATED. *In this volume: Chapter 5*, 2010.
- [6] N. Blasco, P. Corredor, C. Del Rio, and R. Santamaria. Bad news and dow jones make the spanish stocks go round. *European Journal of Operational Research*, 163(1):253 – 275, 2005.
- [7] J.H. Boyd, J. Hu, and R. Jagannathan. The stock market’s reaction to unemployment news: Why bad news is usually good for stocks. *The Journal of Finance*, 60(2):649–672, 2005.
- [8] Brown. *CARISMA annual conference: Incorporating news analytics into quantitative investment and trading strategies* <http://www.optirisk-systems.com/papers/RichardBrown.pdf>, 2010.
- [9] J. A. Busse and T. Clifton Green. Market efficiency in real time. *Journal of Financial Economics*, 65(3):415–437, 2002.
- [10] R. Cahan, J. Jussa, and Y. Luo. Breaking news: How to use news sentiment to pick stocks. *MacQuarie Research Report*.
- [11] J.Y. Campbell, A.W. Lo, and A.C. MacKinlay. The econometrics of financial markets. *Chapter 4: Event study analysis*.
- [12] W.S. Chan. Stock Price Reaction to News and No-News: Drift and Reversal After Headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.
- [13] Z. Da, J. Engelberg, and P. Gao. In Search of Attention. Working Paper: Available on SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1364209, 2009.
- [14] K. Daniel, D. Hirshleifer, and A. Subrahmanyam. Investor psychology and security market under-and overreactions. *The Journal of Finance*, 53(6):1839–1885, 1998.
- [15] K. Daniel, D. Hirshleifer, and S.H. Teoh. Investor psychology in capital markets: Evidence and policy implications. *Journal of Monetary Economics*, 49(1):139–209, 2002.
- [16] S. Das. News analytics metrics: Desirable properties TBC. *In this volume: Chapter 3*, 2010.
- [17] S.R. Das and M.Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.
- [18] W. De Bondt and R. Thaler. Does the stock market overreact? *The Journal of Finance*, 40(3):793–805, 1985.
- [19] D. diBartolomeo. Using news as a state variable in assessment of financial market risk. *In this volume: Chapter 9*, 2010.

- [20] D. diBartolomeo and S. Warrick. Making covariance based portfolio risk models sensitive to the rate at which markets reflect new information. In Knight J. and Satchell. S., editors, *Linear Factor Models*. Elsevier Finance, 2005.
- [21] M. Dzielinski, M.O. Rieger, and T. Talpsepp. Volatility, asymmetry, news and private investors. *In this volume: Chapter 10*, 2010.
- [22] L.H. Ederington and J.H. Lee. How markets process information: News releases and volatility. *Journal of Finance*, 48:1161–1191, 1993.
- [23] J. Engleberg and S. Sankaraguruswamy. How to gather data using a web crawler: An application using SAS to search EDGAR. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1015021&r...
- [24] E.F. Fama and K.R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–466, 1992.
- [25] L.H. Fang and J. Peress. Media coverage and the cross-section of stock returns. *Forthcoming in Journal of Finance*, 2009.
- [26] M. Graham, J. Nikkinen, and P. Sahlstrom. Relative Importance of Scheduled Macroeconomic News for Stock Market Investors. *Journal of Economics and Finance*, 27(2):153–165, 2003.
- [27] R.C. Gutierrez Jr, E. Kelley, and M.C. Hall. The long-lasting momentum in weekly returns. *Journal of Finance*, Forthcoming.
- [28] Hafez. *CARISMA annual conference: The role of news in financial markets* <http://www.optirisk-systems.com/papers/PeterAgerHafez.pdf>, 2010.
- [29] P. Hafez. Detection of seasonality in newsflow. *White Paper available from RavenPack*, 2009.
- [30] D. Hirshleifer. Investor psychology and asset pricing. *The Journal of Finance*, 56(4):1533–1597, 2001.
- [31] D. Hirshleifer, S.S. Lim, and S.H. Teoh. Driven to Distraction: Extraneous Events and Underreaction to Earnings News (Digest Summary). *CFA Digest*, 40(1), 2010.
- [32] H. Hong and J.C. Stein. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184, 1999.
- [33] K. Hou, L. Peng, and W. Xiong. A tale of two anomalies: The implications of investor attention for price and earnings momentum. Available <http://ssrn.com/abstract=976394>.
- [34] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.
- [35] P.S. Kalev and H.N. Duong. Firm specific news arrival and the volatility of intraday stock index and futures returns. *In this volume: Chapter 11*, 2010.
- [36] J.M. Karpoff. The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis*, 22:109–126, 1987.
- [37] J. Kittrell. Sentiment reversals as buy signals. *In this volume: Chapter 8*, 2010.
- [38] SP Kothari and J.B. Warner. Econometrics of event studies. In *Handbook of Empirical Corporate Finance*. Elsevier Finance, 2005.
- [39] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management*. ACM, 2000.
- [40] D. Leinweber. *Nerds on Wall Street*. John Wiley, 2009.
- [41] D. Leinweber and J. Sisk. Relating news analytics to stock returns. *In this volume: Chapter 4*, 2010.

- [42] F. Li. Do stock market investors understand the risk sentiment of corporate annual reports? *University of Michigan Working Paper Available http://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181*.
- [43] F. Li. Do stock market investors understand the risk sentiment of corporate annual reports? UPDATED. *In this volume: Chapter 6*, 2010.
- [44] X. Liang. Impacts of internet stock news on stock markets based on neural networks. In *Advances in neural networks*. Springer Berlin/Heidelberg, 2005.
- [45] J. Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, 1965.
- [46] A. Lo. Reuters NewsScope Event Indices. *AlphaSimplex research report produced in partnership with Thomson Reuters*.
- [47] A. Lo. *Market efficiency: Stock market behaviour in theory and practice*. Edward Elgar Pub, 1997.
- [48] A. Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 2004.
- [49] A. Lo and A. Healy. Reuters NewsScope Event Indices. *In this volume: Chapter 2*, 2010.
- [50] T. Loughran and B. McDonald. When is a liability not a liability? *Journal of Finance*, Forthcoming, 2010.
- [51] R. Luss and A. d’Aspremont. Predicting abnormal returns from news using text classification. *Working Paper from ORFE Princeton*, 2009.
- [52] R.C. Merton. A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, pages 483–510, 1987.
- [53] L. Mitra, G. Mitra, and D. diBartolomeo. Equity portfolio risk (volatility) estimation using market information and sentiment. *Quantitative Finance*, 9(8):887–895, 2009.
- [54] M.A. Mittermayer and G. Knolmayer. Text mining systems for market response to news: A survey. *Working Paper University of Bern: Available on SSRN <http://www2.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-184.pdf>*, 2006.
- [55] A. Moniz, G. Brar, and C. Davis. Have I Got News for You. *MacQuarie Research Report*.
- [56] A. Moniz, G. Brar, and C. Davis. Have I Got News for You. *In this volume: Chapter 7*, 2010.
- [57] Munz. *CARISMA annual conference: US markets: Earnings news release - an inside look <http://www.optirisk-systems.com/papers/MarianMunz.pdf>*, 2010.
- [58] T. Odean and B. Barber. Are investors reluctant to realize their losses? *The Journal of Finance*, 53(5):1775–1798, 1998.
- [59] A. Patton and M. Verardo. Does Beta Move with News? Systematic Risk and Firm-Specific Information Flows. *FMG Discussion Papers available from <http://eprints.lse.ac.uk/24421/1/dp630.pdf>*, 2009.
- [60] D. Peramunetilleke and R. K. Wong. Currency exchange rate forecasting from news headlines. In Xiaofang Zhou, editor, *Thirteenth Australasian Database Conference (ADC2002)*, Melbourne, Australia, 2002. ACS.
- [61] R.L. Peterson. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioral Finance*, 8(2):70–78, 2007.
- [62] Kalev P.S., W.M Liu, P.K. Pham, and E. Jarneic. Public information arrival and volatility of intraday stock returns. *Journal of Banking and Finance*, 28(6):1441–1467, 2004.
- [63] RavenPack. RavenPack News Scores User Guide. *February 11, 2010, Version 1.3.1*, 2010.

- [64] C. Robertson, S. Geva, and R. Wolff. What types of events provide the strongest evidence that the stock market is affected by company specific news? In *Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61*, page 153. Australian Computer Society, Inc., 2006.
- [65] C.S. Robertson, S. Geva, and R.C. Wolff. News aware volatility forecasting: Is the content of news important? In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 161–170. Australian Computer Society, Inc., 2007.
- [66] B. Rosenberg, K. Reid, and R. Lanstein. Persuasive evidence of market inefficiency. *Journal of Portfolio Management*, 11(3):9–16, 1985.
- [67] S.A. Ross. The Arbitrage Pricing Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.
- [68] P. Ryan and R.J. Taffler. Are economically significant stock returns and trading volumes driven by firm-specific news releases? *Journal of Business Finance & Accounting*, 31(1-2):49–82, 2004.
- [69] I. Schmerken. Trading off the news. *Wall Street and Technology*, Available from http://www.wallstreetandtech.com/technology-risk-management/showArticle.jhtml;jsessionid=ZYNMF1D4EJ4LHQE1GHRSKHWATMY32JVN?articleID=185302817&_requestid=532279.
- [70] J. Scott, M. Stumpp, and P. Xu. News, not trading volume, builds momentum. *Financial Analysts Journal*, 59(2):45–54, 2003.
- [71] M. Seasholes and G. Wu. Profiting from predictability: Smart traders, daily price limits, and investor attention. *University of California, Berkeley, working paper available <http://www.nd.edu/~pschultz/SeasholesWu.pdf>*, 2004.
- [72] W.F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance*, 19(3):425–442, 1964.
- [73] A Team. Machine readable news and algorithmic trading. *Whitepaper produced for Thomson Reuters and Market News International*, 2010.
- [74] P.C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168, 2007.
- [75] P.C. Tetlock, M. Saar-Tsechansky, and S. Mackassy. More than words: Quantifying language to measure firms’ fundamentals. *Journal of Finance*, 63(3):1437–1467, 2008.
- [76] Thomson-Reuters. Reuters NewsScope Sentiment Engine: Guide to Sample Data and System Overview. *Dec, 2009*, Version 3, 2009.
- [77] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453, 1981.
- [78] Vreijling. *CARISMA annual conference: Practical use of news in equity trading strategies <http://www.optirisk-systems.com/papers/MarkVreijling.pdf>*, 2010.
- [79] Wysocki. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. *Working Paper No. 98025 University of Michigan*, 1999.

A Appendix I: Structure and Content of News Data

In this appendix we summarise the services by two leading providers of news analytics, namely Thomson Reuters (NewsScope) and RavenPack (NewsScore). The information is presented under three main headings, (i) Coverage, (ii) Method and Types of Cores, (iii) Example of News Data in a tabular form.

Details of Thomson Reuters News Analytics Coverage

(i) Coverage

Real-time and historical equity coverage

Commodities and Energy: 39 C&E Topics

Equity:

All Equities	34,037	100.00%
Active companies	32,719	96.1%
Inactive companies	1,318	3.9%

Equity Coverage by region

Americas: 14,785

APAC: 11,055

EMEA: 8,197

Equity coverage updates: Bi-Weekly upgrade to recent changes

History: Available from January, 2003 (history kept for de-listed companies; symbology change tracked).

Data Fields: 82 metadata fields including Timestamp (GMT Millisec), linked counts over various time periods which measure repetition, linked item cross references, language, topics, prevailing sentiment, detailed sentiment, relevance, size of item, broker action, market commentary, number of companies mentioned, position of first mention, news intensity, news source, story type, headline, company identifier, among others.

Delivery Mechanisms: Internet/VPN, co-lo, dedicated circuits, deployed on-site, FTP, Thomson Reuters Quantitative Analytics / Market QA.

News Sources: Reuters host of third party sources standard; Able to process customer specific sources including internet feeds, PDF files, and text from databases.

(ii) Method and Types of Scores

(Thomson Reuters : Please supply a short description of column headings)

(iii) Example of News Data in a tabular form

TIMESTAMP	R/C	RELEVANCE	SENTIMENT	POSITIVE	NEUTRAL	NEGATIVE	LINKED COUNTS	ITEMTYPE	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 SPWR 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 - THEFLYONTHEWALL.COM	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 SPWR 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. AFTER	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
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 SPWR 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 - THEFLYONTHEWALL.COM	RCH US DPR HDWR SPWR 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.com	RCH US DPR HDWR SPWR 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 SPWR 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 FROM	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 9: Thomson Reuters NewsScope Sentiments

Details of RavenPack NewsScores: Equity coverage and available data

(i) Coverage

Real-time and historical equity coverage

Equity coverage	10,742	100.00%
Active companies	9,934	92.48%
Inactive companies	808	7.42%

Coverage by region

Americas:	4,141
Asia:	3,642
Europe:	2,431
Oceania:	353
Africa:	175

Available data

Historical:

Data format:	Comma separated values (CSV) file
Archive range:	Single .csv files comprising coverage of all companies in their same sector 41 sector files plus 1 uncategorised = 42 .csv files compressed in .zip files on a per year basis
Data fields:	10 fields: Date/time, ID, Sentiment (5), Market Impact, Category and Company relevance
Download:	Secure web download

Real-time:

Connection:	Over the internet
Software:	Local installation of RavenPack Data Gateway(Windows Client) plus API
Access:	Push feed for real-time plus historical query mechanism to fill gaps as required
API:	Java

(ii) Method and Types of Scores

TIMESTAMP_UTC The Date/Time (yyyy-mm-dd hh:mm:ss.000) at which the news item was published in Coordinated Universal Time (UTC).

COMPANY_IDENTIFIER A unique and permanent company identifier assigned by RavenPack that consistently identifies companies throughout the historical archive and in real-time. Company identifiers are mapped to common securities identifiers such as ISINs, CUSIPs, TICKERs, etc.

COMPANY_RELEVANCE A score between 0-100 that indicates how strongly related the company is to the underlying news story, with higher values indicating greater relevance.

EVENT_CATEGORIES An element or "tag" representing a company-specific news announcement or formal event. Highly relevant stories about companies are classified into a set of predefined event categories. When applicable, the role played by the company in the story is also detected and tagged.

EVENT_SENTIMENT A score between 0 and 100 that represents the news sentiment for a given company by measuring various sentiment proxies sampled from the news. The score is determined by systematically matching stories typically categorized by financial experts as having short-term positive or negative share price impact. The algorithm produces a score for more than 160 types of news events - scheduled and

unscheduled.

EVENT _NOVELTY A score between 0 and 100 that represents how "new" or novel a news story is within a given time window.

EVENT _NOVELTY _KEY An identifier that provides a way to chain or relate similar stories about a similar company-specific event. Its a powerful way to build a model based on relationships between companies and event categories like product recalls, layoffs, corporate or legal issues, earnings announcements, analyst and credit revisions, and many more.

COMPOSITE _SENTIMENT A sentiment score between 0 and 100 that represents the news sentiment of a given story by combining various sentiment analysis techniques. The direction of the score is determined by looking at emotionally charged words and phrases and by matching stories typically rated by experts as having short-term positive or negative share price impact. The strength of the score (values above or below 50, where 50 represents neutral strength) is determined from intraday stock price reactions modeled empirically using tick data.

STORY LEVEL _SENTIMENT A collection of sentiment scores between 0 and 100 that represent the sentiment of a given story. RavenPack includes various metrics, each calculated using a different linguistic classifier or technique. A classifier is an algorithm designed to assess the overall sentiment of a company by detecting events and language within stories that are likely to drive stock prices upwards or downwards over a given period of time.

(iii) Example of News Data in a tabular form

TIMESTAMP UTC	RP STORY ID	WLE	PCM	ECM	RCM	VCM	NIP	COMPANIES
2009-05-13 18:55:33	761C6BB4D362C6106FC51A5DA7F39074	50	0	50	50	50	41	US/IHS:2
2009-05-13 18:55:44	C5281C388E3AFE07D6AB866216034AE5	50	50	100	50	50	27	US/RAI:90
2009-05-13 18:55:56	34501433353BD6A9BF1DD6F0D6C5CC47	50	50	0	50	50	41	US/BAC:90
2009-05-13 18:56:03	B62C859C48AE29FA81586CC95A764B33	50	50	100	100	50	42	US/IBM:100
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BAC:12,GB/BARC:2,US/RF:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BAC:12,GB/BARC:2,US/RF:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/TOL:2,US/DHI:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/TOL:2,US/DHI:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BA:12

Source: RavenPack, Macquarie Capital (USA), May 2009

Figure 10: RavenPack NewsScores

A.1 Appendix II: Annotated bibliography of selected papers

A.1.1 Reuters NewsScope Event Indices (NEI)

Author:

Andrew Lo, AlphaSimplex

Focus:

The creation of event indices (Reuters NewsScope Event Indices NEI) which reflect issuance of market moving news. The indices are constructed to have “predictive” power for (realised) volatility and returns, since they are constructed in an integrated framework where news, returns and (realised) volatility are considered in creating the indices. The indices are designed to form inputs into systematic investment and risk management protocols.

Method:

The framework for developing the Reuters NEI is as follows. For a given asset class and related topic area the following parameters are used.

- (1) List of keywords and phrases with real valued weights; $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$.
- (2) A rolling “sentiment” window of size r (say 5/10 minutes).
- (3) A rolling calibration window of size R (say 90 days).

Initially a *raw score* is created.

We have $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$, where W_1 is the first keyword and γ_1 is the weighting for the first keyword.

The raw score at time t is assigned by considering the “sentiment” window $(t - r, t]$. (w_1, \dots, w_k) is the vector of keyword frequencies in $(t - r, t]$, that is, w_i is the number of times keyword W_i occurred in the last r minutes. The raw score is defined as

$$s_t \equiv \sum_i \gamma_i w_i \quad (4)$$

The raw score will tend to be high when the news volume is high. A *normalised score* is therefore produced using the rolling calibration window. At all times t for the R days in the calibration window, we record

- (i) the raw score s_t that would have been assigned,
- (ii) the news volume; $n_{[t-r, t]}$ the number of words that were observed in the time interval $[t - r, t)$.

The normalised score is determined by comparing the current raw score against the distribution of raw scores in the calibration window, where the news volume equalled the current news volume. This means we only consider those raw scores where the news volume equals the current news volume.

$$S_t \equiv \frac{|\{t' \in [t - R, t) : n_{[t'-r, t']} = n_{[t-r, t]} \ \& \ s_{t'} < s_t\}|}{|\{t' \in [t - R, t) : n_{[t'-r, t']} = n_{[t-r, t]}\}|} \quad (5)$$

The numerator is a subset of the denominator, hence $S_t \leq 1$. If $S_t = 0.92$, we can say 92% of the time when news volume is at the current level, the raw score is less than it currently is. Lo creates an alternative score based on topic codes. Instead of counting word frequencies, the fraction of news alerts (in the last r

minutes) tagged with particular topic codes, are used.

Naturally the scoring method is dependent on the list of keywords/topic areas (W_1, \dots, W_k) and the real valued weights ($\gamma_1, \dots, \gamma_k$). The lists of keywords/topics were created by selecting the major news categories that related to the asset class (foreign exchange) and creating lists, by hand, of words and topic areas that suggest news relevant to the categories. A tool was created to extract news from periods where high scores were assigned. This news was then manually inspected, so that the developer could determine whether the keywords (topics) were legitimate or needed adjusting.

The optimal weights ($\gamma_1, \dots, \gamma_k$) for the intraday return sentiment index were determined by regressing the keyword/topic frequencies against the intraday asset returns. Similarly the (optimal) weights for the intraday volatility sentiment index were determined by regressing the keyword/topic frequencies against the intraday (de-seasonalised) realised volatility. Volatility was observed to show strong seasonality on intraday timescales, hence this series was de-seasonalised prior to derivation of the weights. Returns did not exhibit any seasonality. The time series are given on an intraday basis, hence to keep the data manageable a random subset of the observations are used in calibration. Lo notes the determination of the weights can be expressed as a more general classification problem. Other techniques might be applied, in particular machine learning algorithms such as the perceptron algorithm or support vector machines. He suggests further study is required to find the best approach, but the standard linear regression approach does perform well.

Data source:

Reuters NewsScope Alerts which are news flashes issued when newsworthy events occur. These items are both timely and relevant. They are tagged with machine readable codes. The alerts' text is concise and formed from a relatively small vocabulary, hence lends itself well to applications of machine learning algorithms.

Results and conclusions:

Lo undertakes detailed event study analysis to establish that the final NEI have empirical significance. He uses the NEI series to define an event. An event is defined to take place when the index exceeds a certain threshold (say 0.995). He then removes any events that follow in less than one hour of another event. This guards against identifying "new" events which are actually based on old news. The behaviour of exchange rates before and after these events are then studied. Two time series are considered; the log returns and the deseasonalised squared log returns. He then tests the null hypothesis that the distribution of log returns / deseasonalised squared log returns are the same before and after the events. He uses samples of one hour centered on the events. Lo finds that the event studies confirm the constructed event indices, on average, impact the *realised* foreign exchange volatility.

A.1.2 Yahoo! for Amazon: Sentiment extraction from small talk on the web

Author:

Sanjiv Das and Mike Chen

Focus:

Das and Chen (2007) use statistical and natural language techniques to extract investor sentiment from stock message boards and generate sentiment indices. They apply their method for 24 technology stocks present in the Morgan Stanley High Tech (MSH) Index.

Method:

A web scraper program is used to download tech sector message board messages. Five algorithms, each with different conceptual underpinnings are used to classify each message. A voting scheme is then applied to all five classifiers.

Three supplementary databases are used in the classification algorithms.

1. “*Dictionary*” is used for determining the nature of the word. For example, is it a noun, adjective or adverb?
2. “*Lexicon*” is a collection of hand picked finance words which form the variables for statistical inference within the algorithms.
3. “*Grammar*” is the training corpus of base messages used in determining the in-sample statistical information. This information is then applied for use on the out-of-sample messages.

The lexicon and grammar jointly determine the context of the sentiment. Each of the classifiers relies on a different approach to message interpretation. They are all analytic, hence computationally efficient.

1. *Naive classifier* (NC) is based on a word count of positive and negative connotation words. Each word in the lexicon is identified as being positive, negative or neutral. A parsing algorithm negates words if the context requires it. The net word count of all lexicon matched words is taken. If this value is greater than one, we sign the message as a buy. If the value is less than one the message is a sell. All others are neutral.
2. *Vector distance classifier* Each of the D words in the lexicon is assigned a dimension in vector space. The full lexicon then represents a D -dimensional unit hypercube. Every message can be described as a word vector in this space ($m \in \mathbb{R}^D$). Each hand tagged message in the training corpus (grammar) is converted into a vector G_j (grammar rule). Each (training) message is pre classified as positive, negative or neutral. We note that Das and Chen use the terms Buy/Positive, Sell/Negative and Neutral/Null interchangeably. Each new message is classified by comparison to the cluster of pretrained vectors (grammar rules) and is assigned the same classification as that vector with which it has the smallest angle. This angle gives a measure of closeness.
3. *Discriminant based classification* NC weights all words within the lexicon equally. The discriminant based classification method replaces this simple word count with a weighted word count. The weights are based on a simple discriminant function (Fisher Discriminant Statistic). This function is constructed to determine how well a particular lexicon word discriminates between the different message categories. These categories are { Buy, Sell, Null }. The function is determined using the pre classified messages within the grammar. Each word in a message is assigned a signed value, based on its sign in the lexicon multiplied by the discriminant value. Then as for NC a net word count is taken. If this value is greater than 0.01, we sign the message as a buy. If the value is less than -0.01 the message is a sell. All others are neutral.

4. *Adjective - adverb phrase classifier* is based on the assumption that phrases which use adjectives and adverbs emphasize sentiment and require greater weight. This classifier also uses a word count but uses only those words within phrases containing adjectives and adverbs. A “tagger” extracts noun phrases with adjectives and adverbs. A lexicon is used to determine whether these significant phrases indicate positive or negative sentiment. The net count is again considered to determine whether the message has negative or positive overall sentiment.
5. *Bayesian classifier* is a multi variate application of Bayes Theorem. It uses the probability a particular word falls within a certain classification and is hence indifferent to the structure of language. We consider three categories $C = \{c_i \mid i = 1, \dots, C\}$. Denote each message $m_j \quad j = 1, \dots, M$. The set of lexical words is $F = \{w_k\}_{k=1}^D$. (The total number of lexical words is D) We can determine a count of the number of times each lexical item appears in each message $n(m_j, w_k)$. Given the class of each message in the training set we can determine the frequency with which a lexical word appears in a particular class. We are then able to compute the conditional probability of an incoming message j falling in category i , $Pr(m_j|c_i)$, from the word based frequencies. $Pr(c_i)$ is set to the proportion of messages in the training set classified in class c_i . For a new message we are able to compute the probability it falls within class c_i given its component lexicon words, that is $P(c_i|m_j)$, through an application of Bayes Theorem. The message is classified as being from the category with the highest probability.

A voting scheme is then applied to all five classifiers. The final classification is based on achieving a majority amongst the five classifiers. If there is no majority the message is not classified. This reduces the number of messages classified but enhances the classification accuracy.

Das and Chen also introduce a method to detect message ambiguity. Messages posted on stock message boards are often highly ambiguous. The grammar is often poor and many of the words do not appear in standard dictionaries. [They note “Ambiguity is related to the absence of “aboutness””. The General Inquirer has been developed by Harvard University for content analyses of textual data. They use it to determine an independent optimism score for each message. By using a different definition of sentiment it is ensured there is no bias to a particular algorithm. The optimism score is the difference between the number of optimistic and pessimistic words as a percentage of the total words in the body of the text. This score allows us to rank the relative sentiment of all stories within a classification group. For example, they can rank the relative optimism of all stories which have been classified by their scheme as positive. The mean and standard deviation of the optimism score for different classification types ($\{\text{Buy, Sell, Null}\}$) can be calculated. They filter *in* and consider only highly optimistically scored stories in the positive category. For example only those stories with optimism scores above the mean value plus one standard deviation are considered. Similarly they filter *in* and consider only the most highly pessimistic scores in the negative category.] Once the classified stories are further filtered for ambiguity, it is found that the number of false positives dramatically declines.

Once the sentiment for each message is determined using the voting algorithm, a daily sentiment index is compiled. The classified messages up to 4pm each day are used to create the aggregate daily sentiment for each stock. A buy (sell) message increments (decrements) the index by one. These indices are further aggregated across all stocks to obtain an aggregate sentiment for the technology portfolio.

Data source:

Results and conclusions: