Applications of news analytics in finance: A review

Leela Mitra and Gautam Mitra

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 recognized 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 categorization are emerging. Further machine-learning techniques are 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 are 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.

1.1 INTRODUCTION

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 an 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; Barber and Odean, this volume, Chapter 7; diBartolomeo and Warrick, 2005; Mitra, Mitra, and diBartolomeo, 2009; Dzielinski, Rieger, and Talpsepp, this volume, Chapter 11). 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, Barber and Odean (this volume, Chapter 7) 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 that volume increases on days with information releases (Bamber, Barron, and Stober 1997; Karpoff, 1987; Busse and Green, 2004). Important news frequently results in large positive or negative returns. Ryan and 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 characterized as being non-numeric and hard to quantify. Unlike analysis based on quantified market data, textual news data contain information about the effect of an event and the possible causes of an event. It is natural to expect that the application of these 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 categorize 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 minutiae of repetitive analysis, such that they are able to better target their reading and research. These technologies reduce the burden of 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 the 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 diversi-
fied, 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).

As shown in Figure 1.1 news data are an additional source of information that can be harnessed to enhance (traditional) investment analysis. Yet it is important to recognize 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 needs 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 organized as follows. In Section 1.2 we consider the different sources of news and information flows which can be applied for updating (quantitative) investor beliefs and knowledge. Section 1.2.2 covers several aspects of pre-analysis to be considered when using news in trading systems and quantitative models. In Section 1.3 we consider how qualitative text can be converted to quantified metrics which can form inputs to quantitative models. In Section 1.4 we present news-based models; in particular, we consider the computational architecture (Section 1.4.1), applications for trading and fund management (Section 1.4.2) and applications for

---

**Figure 1.1.** A simple representation of news analytics in financial decision making.
risk management (Section 1.4.3). In Section 1.4.4 desirable industry applications are outlined. The summary conclusions are presented in Section 1.5.

1.2 NEWS DATA

1.2.1 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 are also progressively growing in volume and number.

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 macroeconomic 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, the blog of BBC’s Robert Peston). 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 relatively more attention to the second two sources of news than institutional investors (Dzielinski, Rieger, and Talpsepp, this volume, Chapter 11; Das and 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, 2009, Ch. 10). 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, Google Trends is a free but limited service which provides an 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](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. Connote provide services which allows customized automated collection of SEC filing information for customers (fund managers and traders). Engleberg and Sankaraguruswamy (2007) consider how to use a web crawler to mine SEC filing information through EDGAR.

As stated in Section 1.1, financial news can be split into regular synchronous announcements (scheduled or expected news) and event-driven asynchronous announcements (unscheduled or unexpected news). Mainstream news, rumours, and social media normally arrive asynchronously in an unstructured textual form. A substantial portion of pre-news arrives 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 macroeconomic announcements and earnings announcements. Macroeconomic news, particularly economic indicators from the major economies, is widely used in automated trading. It has 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 on 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 capitalize on events is substantial. Content publishers often specialize in a few data items and hence trading firms often multisource 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 stock prices. Scheduled earnings announcement information is also widely anticipated and used within trading strategies. The pace of response to announcements has accelerated greatly in recent years (see Leinweber, 2009, pp. 104–105). 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. They are characterized as being non-numeric and difficult to process quickly and quantitatively. Unlike analysis based on quantified market data, textual news data contain information about the effect of an event and the possible causes of an event. However, to be applied in trading systems and quantitative models they need 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 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 vendors 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 Section 1.A (the appendix on p. 25). In the appendix we have summarized some essential information taken from the RavenPack News Analytics—Dow Jones Edition (RavenPack, 2010) and Thomson Reuters NewsScope Sentiment Engine (Thomson Reuters, 2009).

1.2.2 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 volume of news data available for different companies*. Larger companies (with more liquid stock) tend to have higher news coverage/news flow. Moniz, Brar, and Davis (2009) observe that the top quintile 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 1.2).

*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

![Figure 1.2. Number of news items vs. log market capitalization (taken from Cahan, Jussa, and Luo, 2009).](image)

\[ y = 0.508x - 0.2752 \]

\[ R^2 = 0.2881 \]
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 story’s topic areas and other important metadata. For example, they may include information about which company a story is about. Tagged stories held by major newswire providers are also accurately time-stamped. The SEC is pushing to have companies file their reports using XBRL (eXtensible Business Reporting Language). Rich Site Summary (RSS) feeds (an XML format for web content) allow customized, automated analysis of news events from multiple online sources.

Tagged news stories provide us with hundreds of different types of events, so that we can effectively use these stories. We need to distinguish what types of news are relevant for a given model (application). Further, the market may react differently to different types of news. For example, Moniz, Brar, and Davis (2009) find 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; Campbell, Lo, and MacKinlay, 1996 for in-depth reviews of event study methodology). Leinweber (this volume, Chapter 6) uses Thomson Reuters tagged news data to investigate several news-based event strategies. Elementized 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, Brar, and Davis (2009) find 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 (CARs). They use 350,000 news stories about S&P 500 companies appearing in the Wall Street Journal and Dow Jones News Service from 1984 to 2004. Each story’s (language) sentiment is determined using the General Inquirer and a story is classified as either positive or negative. The CARs for each story classification type relative to the date of the news release are shown in Figure 1.3. There seems to be a connection between a news story’s release and CARs. However, there also seems to be some “information leakage” since CARs 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.

In order to deal with potential noise, Reuters identifies relevance scores for different news articles. Such scores measure how pertinent an article is to a particular company
and helps prevent erroneous links between stories and entities. In particular, after filtering by relevance as measured by RavenPack, Hafez (2009a) obtains a $3\times$ improvement in correlations between a calculated market sentiment measure and out-of-sample returns. Both Reuters and RavenPack include measures for article novelty (uniqueness) which determines repetition among articles and how many similar articles there are for a particular company. In addition, RavenPack (2010) measures event novelty based on more than 200 event categories that are automatically detected in the news. This allows the user to consider not only the first instance of a company event but also to measure how much media attention it receives.

Several studies also report strong seasonality in news flow at hourly, daily and weekly frequencies (Lo, 2008; Hafez, 2009b; Moniz, Brar, and Davis, 2009). A valuable aspect of pre-analysis of news data is to identify periods of unexpected news flow levels, from periods of variation due to seasonality, in order to identify periods where significant levels of information are flowing into the market. Hafez (2009b) investigates the seasonality patterns of news arrival. Figures 1.4 and 1.5 show the intraday pattern. He notes that larger volumes of news flow arrive just before the opening of the European, US, and Asian trading sessions. On the intra-week level we can see little news flow takes place at the weekends. During the week, the peak of news flow occurs on Wednesday and Thursday, while the trough falls on Friday. Lo also notes that the median number of weekday Reuters news alerts is usually between 1,500 and 2,000, while the median for the entire weekend drops to around 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 et al., 2005; Boyd, Hu, and Jagannathan, 2005; Liang, 2005; Tetlock, 2007). We discuss how to extract the informational content of news (that is, the sentiment) in Section 1.3. It has been recognized 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 normalized 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 company-specific news. 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.

The relationship between different news stories 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 its CEO and provide guidance on its sales outlook. The dependence or independence between different news stories is a consideration.

Figure 1.4. Seasonality—intraday pattern.
1.3 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 language training sets developed by finance experts. Further, dictionary-based algorithms which use psychology-based interpretations of words may be used. Since different groups of people are affected by events differently and have different interpretations of the same events, conflicts may arise. Moniz, Brar, and Davis (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 is better positioned to repay its debts. Loughran and McDonald (forthcoming) also consider how context affects interpretation of the tone of text. 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.

Figure 1.5. Seasonality—intraweek pattern.
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 (i) returns and (ii) volatility. So he assumes that sentiment measured in the context of these two variables is different. This approach is quite pragmatic and is focused on using the news content directly in the context that the modeller is interested in. Lavrenko et al. (2000), Moniz, Brar, and Davis (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/SemLab, 2010) provides 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 scores then allow us to develop systematic investment and risk management processes. Linking these sentiment scores to the asset returns, trading volumes and volatility or, in other words, discovering the connection between news analysis and the financial analytics and the financial analytics models is a leading challenge in this domain of application.

The definition of market sentiment is very much context-dependent. In general, we are interested in discovering the “informational content of news”. In this review chapter, for the purpose of (quantitative) modelling applications, we use the 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). The former uses natural language processing and machine learning whereas the latter applies a market-based measure. Both papers cover the following items:

1. A definition of the context of sentiment.
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 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 in-sample statistical information. This information is then applied for use on 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 with the cluster of pre-trained 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 multivariate 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 $C = 3$ $c_i$ $i = 1, \ldots, C$. Denote each message $m_j$ $j = 1, \ldots, M$. 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 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 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 and has been applied 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, we 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 ({Buy; Sell; Null}) can be calculated. They filter in and consider only 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 decline.

After the sentiment for each message is determined using the voting algorithm, a daily sentiment index is compiled. The classified messages up to 4 pm 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

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

(1.1)

$B$ ($S$) 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. In addition, trading volatility and volume of stocks are calculated and message volume recorded. All the time-series are normalized.
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. The authors plot the two against each other and show that 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, 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 (the $t$-statistic of the 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.

The authors also provide a graphical display of the relationship between the sentiment measure, disagreement measure, message volume, trading volume and volatility. Sentiment is inversely related to disagreement. As disagreement increases, 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 (2004). Trading volume and volatility are strongly related to each other.

Lo (2008) develops the Reuters NewsScope Event Indices (NEIs) which are constructed to have “predictive” power for particular asset returns and (realized) 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 a Reuters NewsScope alert

```
TimeStamp 02 AUG 2007 04:44:26.155
Alert Tsunami Warning Issued for Japan’s Western Hokkaido Coast
JP ASIA NEWS DIS LEN RTRS
```

The alert comprises three items (i) TimeStamp, (ii) a short headline, and (iii) tags and metadata. 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 Reuters NEIs. For a given asset class and related topic area the following parameters are used:
(1) List of keywords and phrases with real-valued weights; \((W_1, \gamma_1), \ldots, (W_k, \gamma_k)\).
(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 \((W_1, \gamma_1), \ldots, (W_k, \gamma_k)\), where \(W_1\) is the first keyword and \(\gamma_1\) is the weighting for the first keyword.

The raw score at time \(t\) is assigned by considering the time period \((t-r, t]\). \((w_1, \ldots, w_k)\) 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

\[
s_t = \sum \gamma_i w_i
\]

(1.2)

The raw score will tend to be high when the news volume is high. A normalized 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 \(n_{[t-r, t]}\); that is, the number of words that were observed in the time interval \([t-r, t]\).

The normalized 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.

\[
S_t = \frac{\{t' \in [t-R, t] : n_{[t-r, t']} = n_{[t-r, t]} \& s_{t'} < s_t\}}{\{t' \in [t-R, t] : n_{[t-r, t']} = n_{[t-r, t]}\}}
\]

(1.3)

We notice the numerator is a subset of the denominator, hence \(S_t \leq 1\). If \(S_t = 0.92\), we can say that 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 \((W_1, \ldots, W_k)\) and the real-valued weights \((\gamma_1, \ldots, \gamma_k)\). 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 \((\gamma_1, \ldots, \gamma_k)\) 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-seasonalized) realized volatility. Volatility was observed to show strong seasonality on intraday time-scales, hence this series was de-seasonalized 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 is used in calibration.
Lo notes the determination of 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 NEIs 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 is then studied. Two time-series are considered: the log returns and the deseasonalized squared log returns. He then tests the null hypothesis that the distribution of log returns/deseasonalized squared log returns are the same before and after the events. He uses samples of one hour centred on the events.

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

1. Figure 1.6 shows the averaged volatility event window. The pre-event (averaged) volatility is shown by a bold line, and the post-event (averaged) volatility is shown by a faint line. There is a peak at the centre where there is a significant increase in volatility.

2. Figure 1.7 shows the density function of pre-event samples and post-event samples of deseasonalized 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-event 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 standard deviation. The $\chi^2$ goodness-of-fit test can be used to determine whether the two samples are likely to have come from different distributions.

**Indices and FX implied volatility**

Lo finds that the event studies confirm that the constructed event indices, on average,
impact realized 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 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.

### 1.4 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.

#### 1.4.1 Information flow and computational architecture

News analytics in finance focuses on improving IT-based legacy system applications. These improvements come through research and development directed at 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, asset (stocks, commodities, FX rates, etc.) prices move (market reactions). The professionals digest these items of information and accordingly make trading decisions, investment decisions and recompute their risk exposures.
The information flow and the (semi) automation of the corresponding IS architecture is set out in Figure 1.8. 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 compute relevant metrics. This is consolidated with the market data of prices and together they constitute the classical datamart which feeds 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 and thereby achieve intelligence amplification (Leinweber, 2009).

1.4.2 Trading and fund management

Generally traders and quantitative fund managers seek to identify and exploit asset mispricings before they are corrected 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 attitude/reactions which result due to the market and analyst misreaction to new information. In particular, this can arise due to delayed information diffusion or due to investor 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 and 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 has deteriorated in recent years, possibly due to wider use of such strategies.

Moniz, Brar, and Davis (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 become a proxy for corporate news flow. 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 that news data can be used to reinforce proxies for news already incorporated in models and that a strategy based on earnings momentum reinforced by news flow is found to be effective.

Event studies based on news events can also provide the cue to fund managers to identify potentially underpriced/overpriced stocks (see the discussion in Section 1.2.2).

**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 run, 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; 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 perception about future asset returns is 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: analyst 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.

Multifactor 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, 1985) 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, e.g., 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 investor behavioural biases (Moniz, Brar, and Davis, 2009; Barber and Odean, this volume, Chapter 7; Seasholes and Wu, 2004). There is a commercial fund called MarketPsy which employs strategies that exploit “collective investor misbehaviour” (see [http://marketpsy.com/](http://marketpsy.com/)).

Barber and Odean 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 men-
tioned on the Dow Jones News Service. In contrast, professional managers who are
better equipped to assess a wider range of stocks are less prone to buying attention-
grabbing stocks. In particular, institutional investors, who use computers to manage
their searches, normally specialize 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 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. But this finding supports the investor recognition
hypothesis of Merton (1987). 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 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
(1985) suggest that investors overreact to recent earnings by placing less emphasis on
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 underreact 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, Stumpp, and Xu (2003) con-
clude that price momentum is caused by underreaction 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 Daniel, Hirshleifer, and Subrahmanyam (1998)
view since one would expect investors to overreact to stocks they are paying close
attention to. Gutierrez, Kelley, and Hall (2007) and Hou, Peng, and Xiong (2009) also
investigate the relationship between news (information flows) and momentum.
1.4.3 Monitoring risk and risk control

For effective financial risk control, fund management 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 have 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 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; diBartolomeo and Warrick, 2005). Recent technological developments have enabled the creation of data-mining tools that can interpret live news feeds (see Section 1.3 and 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 can be broken down as change due to an increase in volatility of that stock and change due to an increase in 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 the S&P 500 are also influenced by macroeconomic announcements. Kalev et al. (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 (this volume, Chapter 12). 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, Rieger, and Talpspepp (this volume, Chapter 11) refer to it as volatility asymmetry. Their investigation concludes it is likely to be driven by the overreaction of private investors to bad news. In line with this theory, they find that 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 macroeconomy, 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 its risk management strategy in predicting (unexpected) high volatility (Leinweber, 2009, Ch. 10, p. 237).

As discussed in Section 1.3, 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.

### 1.4.4 Desirable industry applications

Stock picking, trading, and fund management (Section 1.4.2) and risk control (Section 1.4.3) are established application areas in the finance industry and the use of news analytics (NA) is researched to achieve improved performance. 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. The following is an itemized list of possible/desirable applications:

- **Market surveillance**  Responding to the state of the market and taking into consideration the preoccupations of the watchdogs; 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 and insider information-based trading. An innovative application of NA is to spot patterns which capture these.

- **Trader decision support**  News data can aid traders in making decisions. News data signals may confirm traders’ existing analyses or it may cause them to reconsider their analyses.

- **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 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 (see A Team, 2010).

- **News flow algorithms**  It is widely recognized that news flow 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 Volume Weighted Average Price (NVWAP) vs. VWAP on its own may add value for trade execution strategies.

- **Post-trade analysis**  Assist in proving best execution and trader performance through post-trade analysis.

News data are 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 the non-rational biased behaviour of investors. These can then be exploited.

- **Stock-screening tool**  News data can be used to aid stock screening. In particular, sentiment data may be used to guess the directional movement of future returns. Very good news stocks (e.g., top sentiment quintile) might be selected to be held long and very bad news stocks (e.g., 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.

- **Compliance/Market abuse**  News data may allow regulators to identify potential market abuse and insider trading, perhaps by allowing the regulator to identify market reactions prior to relevant public new releases.

### 1.5 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 chapter 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 chapter 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, artificial intelligence (AI), pattern recognition, text mining, information engineering and financial engineering. We believe news analytics will soon become an important area of study within financial analytics.
1.A APPENDIX: STRUCTURE AND CONTENT OF NEWS DATA

In this appendix we summarize the services offered by two leading providers of news analytics, namely Thomson Reuters (News Analytics) and RavenPack (NewsScore). The information is presented under three main headings: (i) coverage, (ii) method and types of scores, and (iii) example of news data in tabular form.

1.A.1 Details of Thomson Reuters News Analytics equity coverage and available data

(i) Coverage

Real-time and historical equity coverage

Commodities and energy: 39 C&E topics
Equity:

All equities 34,037 100.0%
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 updates for recent changes (de-listings, M&A, IPOs).

History: Available from January 2003 (history kept for delisted companies; symbology changes tracked). Version control procedures enable clients to test “as-was” versions as well as system enhancements. With new enhancements and scoring logic, history is rescored from 2003 and provided to client.

Data fields: Eighty-two metadata fields including Timestamp (GMT in milliseconds), 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-location, dedicated circuits, FTP, Thomson Reuters Quantitative Analytics/market quantitative analytics, and deployed onsite for customers wanting custom analytics and proprietary sources analysed.

News sources: Reuters and a host of third-party sources as standard; able to process customer-specific sources including internet feeds, PDF files, and text from databases.
Method and types of scores

Timestamp (GMT): DD MM YYYY hh:mm:ss The date and time of the news item as time-stamped by the network and written to the News Archive. All messages are time-stamped to the nearest millisecond—this time represents the time that the message was transmitted by Reuters across its real-time network.

Item ID: A unique ID, identifying the news item. If a particular news item is scored for multiple assets (companies or commodities and energy topics), it has the same ID in each of the assets’ metadata sets.

Stock RIC: Reuters Instrument Code (RIC) of the equity (or topic code for commodity and energy items) for which the scores apply. Note: because the system’s sophistication allows for the scoring of items at the individual entity level, not the overall article level sentiment which tends to be less accurate for specific entities, a single news article may produce multiple “rows” or images of data corresponding to each Stock RIC (or C&E topic) in the article.

Feed ID: Feed identifier: The identifier for the feed handler service that supplied the news item. Consists of the feed type, followed by feed service. Useful in determining source or feed credibility and patterns for and effects of news syndication.

News source: This identifies the publisher of the news item within the feed. For example, the originator of a news story published widely on the internet. It is up to the feed handler to supply a value.

Headline: The headline of the news item. For Thomson Reuters, if the news item was an alert, this is the text of the alert. If it was an article, append or overwrite, then this is the headline.

Relevance: A real-valued number indicating the relevance of the news item to the asset. It is calculated by comparing the relative number of occurrences of the asset with the number of occurrences of other organizations and/or commodities within the text of the item. For stories with multiple assets, the asset with the most mentions will have the highest relevance. An asset with a lower number of mentions will have a lower relevance score.

Number of sent wds/tkns: Number of sentiment words/tokens: The number of lexical tokens (words and punctuation) in the sections of the item text that are deemed relevant to the asset. This is the number of words used in the sentiment calculation for this asset. Can be used in conjunction with Total Wds/Tkns to determine the proportion of the news item discussing the asset.

Total wds/tkns: The total number of lexical tokens (words and punctuation) in the item. Can be used in conjunction with Number of Sent Wds/Tkns to determine the proportion of the news item discussing the asset.

First mention: The first sentence in which the scored asset is mentioned. Often, more relevant assets are mentioned towards the beginning of a news item. Can be used in conjunction with Total Sentences to determine the relative position of the first mention in the item.
Total sentences: The total number of sentences in the news item. Can be used in conjunction with First Mention to determine the relative position of the first mention of the asset in the item.

Number of companies: The number of companies in the news item. The CO_IDS field contains a list of company RICs for scoring and is assigned by the feed handler. It is useful to determine if this asset is one of many discussed in the news item (e.g., a round-up article).

Sentiment classification: This field indicates the predominant sentiment class for a news item with respect to this asset. The indicated class is the one with the highest probability. Values are $1 = \text{positive}; 0 = \text{neutral}; -1 = \text{negative}$. Scores are assigned to specific entities (or commodity topics) within the news item.

POS: Positive Sentiment Probability: The probability that the sentiment of the news item was positive for the asset. Range 0–1.0. The three probabilities (POS, NEU, NEG) sum to 1.0. Probability scores are assigned to specific entities (or commodity topics) within the news item.

NEU: Neutral Sentiment Probability: The probability that the sentiment of the news item was neutral for the asset. Range 0–1.0. The three probabilities (POS, NEU, NEG) sum to 1.0. Probability scores are assigned to specific entities (or commodity topics) within the news item.

NEG: Negative Sentiment Probability: The probability that the sentiment of the news item was negative for the asset. Range 0–1.0. The three probabilities (POS, NEU, NEG) sum to 1.0. Probability scores are assigned to specific entities (or commodity topics) within the news item.

Novelty fields (30 in total): Thomson Reuters News Analytics calculates the novelty of the content within a news item by comparing it with a cache of previous news items that contain the current asset. The comparison between items is done using a linguistic fingerprint, and if the news items are similar for that given asset, they are termed as being “linked”. There are five history periods that are used in the comparison, by default they are 12 hours, 24 hours, 3 days, 5 days, and 7 days prior to the news item’s Timestamp. Customers with deployed solutions can set their own historical look-back period lengths.

Two sets of scores are given:

- **Within feed novelty** News items are only compared with previous items from the same feed.
- **Across feed novelty** News items are compared across all feeds attached to the system.

Each set of scores contain the following fields:

LNKD_CNTn: The count of linked articles in a particular time period gives a measure of the novelty of the news being reported—the higher the linked count value, the less novel the story is for the given asset. If the count is zero, then the current item can be considered novel as there are no similar items reporting the story within the history period.
** LNKD_IDn, LNKD_IDPVn: ** The ITEM_IDS of the five most recent and five oldest linked articles for the longest of the history periods. This can be used to cluster similar items. The Across Feed Novelty identifiers are prefixed with an “X”.

**Volume fields (10 in total):** Thomson Reuters News Analytics calculates the volume news for each asset. A cache of previous news items is maintained and the number of news items that mention the asset within each of five history periods is calculated. By default, the history periods are 12 hours, 24 hours, 3 days, 5 days, and 7 days prior to the news item’s timestamp and are the same as used in the novelty calculations. Thus direct comparisons between similar and total items within the history periods can be achieved.

Two sets of scores are given:

- **Within feed volume**  Volume of news items mentioning the asset within the same feed.
- **Across feed volume**  Volume of news items mentioning the asset across all feeds.

Each set of scores contain the following fields:

**ITEM_CNTn:** The total count of items within the corresponding history period. The across feed volume identifiers are prefixed with an “X”.

**Item genre:** Contains the descriptive of the story genre such as an imbalance message or Reuters news headline tags for the item (e.g., INTERVIEWS, EXCLUSIVES, WRAPUPS, DEALTALK, etc.).

**Broker action:** Item is reporting the action of a broker in their recommendation of the asset. For example, “Goldman upgrades Microsoft to buy from sell” would contain “UPGRADE” in the Microsoft record.

**Commentary:** Indicator that the item is discussing general market conditions, such as after-the-bell summaries. May be used to filter/weight news which describes the stock price, something we may already know by consuming a pricing feed.

**Product permission code:** Permission codes that apply to the record. Thomson Reuters News Analytics currently has two permission levels: LIVE and ARCHIVE. These codes are used to specify what we are allowed to do with the record. LIVE allows usage of data in real time; for example, in an algorithmic trading system. ARCHIVE allows usage for algorithm development and training.

**Item type:** Indicates the type of news item. The following values are possible:

- **Alert**  The news item was generated as a result of an alert. It consists of a single line of text generally written to report a single fact quickly.
- **Article**  Indicates that the news item was a fresh story. The item consists of a headline and body/story text.
- **Append**  The news item was generated by appending text to an existing story take. The news item consists of a headline and story body, where News Analytics scores the entire body of the text, not just the appended section.
- **Overwrite**  The news item was generated by replacing the entire body text of a news story. It consists of a headline and body where the body is the new version of the body.

**Primary news access code (PNAC):** A story identifier used to understand the progression of an event’s coverage. The various parts of a story chain (Alert, Article,
updates, corrections) share the same PNAC (as well as story date/time). It can be used to see which article follows a set of alerts, to which article the update applies, etc.

**Story date/time**: Date/time the first alert or take in the story chain was filed (in GMT)—therefore, STORY_DATE/TIME is the same for all messages in a story chain.

**Take date/time**: The date/time of the news item (i.e., when that particular version/take of the story was published). Story date/time will be consistent across the various takes of the story, but take date/time will be specific to that item.

**Take sequence number**: Sequence number of this alert/take in this story—set to 1 for the first and incremented by 1 for each subsequent alert/take in the story. This can help determine if the item is the second alert or third update to a story, for example.

**Attribution**: Organizational source of the story (i.e., Reuters, PR Newswire, etc.).

**Topic codes**: Topic code(s) for the story which annotate what the story is about. For example, corporate results (RES), mergers and acquisitions (MRG), research (RCH), etc.

**Company codes**: InstrumentRIC(s) for the story (i.e., the symbology for those companies mentioned in the story).

(iii) **Example of news data in tabular form**

Subset of available fields (see Figure 1.9).

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>CNT</th>
<th>ATTRIB</th>
<th>TIME-TYPE</th>
<th>ITEM-GENRE</th>
<th>HEADLINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE NOT DEFINED Wells Fargo completes Wachovia purchase</td>
<td>0.11</td>
<td>1.06</td>
<td>0.47</td>
<td>0.17</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE NOT DEFINED US Treasury to view CB-style rescue case-by-case</td>
<td>0.05</td>
<td>-1.19</td>
<td>0.42</td>
<td>0.09</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE INTERVIEW INTERVIEW Philip side seeks underwriters for bond issue</td>
<td>0.35</td>
<td>-1.39</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTIC</td>
<td>L NOT DEFINED DEUTSCHE BANK CUTS CITIGROUP &lt;50% 2010 SHR VIEW BY S&amp;;P 500</td>
<td>1.00</td>
<td>-1.00</td>
<td>0.13</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE DEALTALK DEALTALK - Universal banks warming battle for hedge fund business</td>
<td>0.13</td>
<td>0.43</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE ALERT NOT DEFINED BERNSTEIN CUTS CITIGROUP INC &lt;CNP&gt; PRICE TARGET TO $55-60</td>
<td>1.00</td>
<td>1.00</td>
<td>0.11</td>
<td>0.65</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE HEADLINE E HEADLINE: STOCKS: Stock Market News</td>
<td>0.35</td>
<td>0.92</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTICLE HEADLINE E HEADLINE: STUKE SPREADS: S&amp;P 500: Needing up, down off on Wall Street</td>
<td>0.20</td>
<td>-1.00</td>
<td>0.13</td>
<td>0.70</td>
</tr>
<tr>
<td>0035090918:47:49</td>
<td>CNT RTRS ARTIC</td>
<td>L NOT DEFINED RECENT F</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.9. Thomson Reuters NewsScope sentiments.
1.A.2 Details of RavenPack News Analytics—Dow Jones Edition: Equity coverage and available data

(i) Coverage

Real-time and historical equity coverage

<table>
<thead>
<tr>
<th>Coverage</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity coverage</td>
<td>28,301</td>
<td>(100.0%)</td>
</tr>
<tr>
<td>Active companies</td>
<td>22,172</td>
<td>(78.3%)</td>
</tr>
<tr>
<td>Inactive companies</td>
<td>6,129</td>
<td>(21.7%)</td>
</tr>
</tbody>
</table>

Coverage by region

<table>
<thead>
<tr>
<th>Region</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Americas</td>
<td>11,950</td>
<td>(42.2%)</td>
</tr>
<tr>
<td>Asia</td>
<td>8,858</td>
<td>(31.3%)</td>
</tr>
<tr>
<td>Europe</td>
<td>5,859</td>
<td>(20.7%)</td>
</tr>
<tr>
<td>Oceania</td>
<td>1,436</td>
<td>(5.1%)</td>
</tr>
<tr>
<td>Africa</td>
<td>186</td>
<td>(0.7%)</td>
</tr>
</tbody>
</table>

Available data

Historical

<table>
<thead>
<tr>
<th>Data format</th>
<th>Comma separated values (.csv) files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive range</td>
<td>Available since January 1, 2000</td>
</tr>
<tr>
<td>Archive packaging</td>
<td>All .csv files compressed in .zip on a per year basis</td>
</tr>
<tr>
<td>Data fields</td>
<td>Company analytics: 23 fields including sentiment, novelty, relevance and categories, among others</td>
</tr>
<tr>
<td>Story coding</td>
<td>11 fields with coding information from the original story</td>
</tr>
<tr>
<td>Download</td>
<td>Secure web download</td>
</tr>
</tbody>
</table>

Real-time

<table>
<thead>
<tr>
<th>Connection</th>
<th>Over the internet or direct leased lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>Local install of RavenPack Data Gateway + API</td>
</tr>
<tr>
<td>Access</td>
<td>Push feed for real-time + historical query mechanism</td>
</tr>
<tr>
<td>API</td>
<td>For Windows and Linux.</td>
</tr>
</tbody>
</table>

(ii) Method and types of scores

TIMESTAMP_UTC: The date/time (yyyy-mm-dd hh:mm:ss.sss) at which the news item was received by RavenPack servers in Coordinated Universal Time (UTC).

COMPANY: This field includes a company identifier in the format ISO_CODE/TICKER. The ISO_CODE is based on the company’s original country of incorporation and TICKER on a local exchange ticker or symbol. If the company detected is a privately held company, there will be no ISO_CODE/TICKER information, only an RP_COMPANY_ID.
**ISIN:** An International Securities Identification Number (ISIN) to identify the company referenced in a story. The ISINs used are accurate at the time of story publication. Only one ISIN is used to identify a company, regardless of the number of securities traded for any particular company. The ISIN used will be the primary ISIN for the company at the time of the story.

**RP_COMPANY_ID:** A unique and permanent company identifier assigned by RavenPack. Every company tracked is assigned a unique identifier comprised of six alphanumeric characters. The RP_COMPANY_ID field consistently identifies companies throughout the historical archive. RavenPack’s company detection algorithms find only references to companies by information that is accurate at the time of story publication (point-in-time sensitive).

**RELEVANCE:** A score between 0 and 100 that indicates how strongly related the company is to the underlying news story, with higher values indicating greater relevance. For any news story that mentions a company, RavenPack provides a relevance score. A score of 0 means the company was passively mentioned while a score of 100 means the company was predominant in the news story. Values above 75 are considered significantly relevant. Specifically, a value of 100 indicates that the company identified plays a key role in the news story and is considered highly relevant (context aware). The classifier detecting companies has access to information about each company including short names, long names, abbreviations, security identifiers, subsidiary information, and up-to-date corporate action data. This allows for “point-in-time” detection of companies in the text.

**CATEGORIES:** An element or “tag” representing a company-specific news announcement or formal event. Relevant stories about companies are classified in a set of predefined event categories following the RavenPack taxonomy. When applicable, the role played by the company in the story is also detected and tagged. RavenPack automatically detects key news events and identifies the role played by the company. Both the topic and the company’s role in the news story are tagged and categorized. For example, in a news story with the headline “IBM Completes Acquisition of Telelogic AB” the category field includes the tag acquisition-acquirer (since IBM is involved in an acquisition and is the acquirer company). Telelogic would receive the tag acquisition-acquiree in its corresponding record since the company is also involved in the acquisition but as the acquired company. Similarly, a story published as “Xerox Sues Google Over Search-Query Patents” is categorized as a patent-infringement. Xerox receives the tag patent-infringement-plaintiff while Google gets patent-infringement-defendant. By definition, a company linked to a category given its role receives a RELEVANCE score of 100.

**ESS—EVENT SENTIMENT SCORE:** A granular score between 0 and 100 that represents the news sentiment for a given company by measuring various 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 strength of the score is derived from training sets where financial experts classified company-specific events and agreed these events generally convey positive or negative sentiment and to what degree. Their ratings are encapsulated in an algorithm.
that generates a score range between 0 and 100 where higher values indicate more positive sentiment while values below 50 show negative sentiment.

**ENS—EVENT NOVELTY SCORE:** A score between 0 and 100 that represents how “new” or novel a news story is within a 24-hour time window. The first story reporting a categorized event about one or more companies is considered to be the most novel and receives a score of 100. Subsequent stories within the 24-hour time window about the same event for the same companies receive lower scores.

**ENS_KEY—EVENT NOVELTY KEY:** An alphanumeric identifier that provides a way to chain or relate stories about the same categorized event for the same companies. The ENS_KEY corresponds to the RP_STORY_ID of the first news story in the sequence of similar events. The identifier allows a user to track similar stories reporting on the same event about the same companies.

**CSS—COMPOSITE SENTIMENT SCORE:** 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 from approximately 100 large-cap stocks.

**NIP—NEWS IMPACT PROJECTIONS:** A score taking values between 0 and 100 that represents the degree of impact a news flash has on the market over the following 2-hour period. The training set for this classifier used tick data for a test set of large-cap companies and looked at the relative volatility of each stock price measured in the 2 hours following a news flash. This NIP score is based on RavenPack’s Market Response Methodology.

**PEQ—GLOBAL EQUITIES:** A score that represents the news sentiment of the given news item according to the PEQ classifier, which specializes in identifying positive and negative words and phrases in articles about global equities. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Traditional Methodology.

**BEE—EARNINGS EVALUATIONS:** A score that represents the news sentiment of the given story according to the BEE classifier, which specializes in news stories about earnings evaluations. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Expert Consensus Methodology.

**BMQ—EDITORIALS & COMMENTARY:** A score that represents the news sentiment of the given story according to the BMQ classifier, which specializes in short commentary and editorials on global equity markets. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Expert Consensus Methodology.
BAM—VENTURE, COMPANY, MERGERS & ACQUISITIONS: A score that represents the news sentiment of the given story according to the BAM classifier, which specializes in news stories about mergers, acquisitions and takeovers. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Expert Consensus Methodology and has been trained on stories that lead up to a pre-identified mergers, acquisitions and takeover event.

BCA—REPORTS ON CORPORATE ACTIONS: A score that represents the news sentiment of the given news story according to the BCA classifier, which specializes in reports on corporate action announcements. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Expert Consensus Methodology and has been trained on stories that lead up to a pre-identified corporate action announcement.

BER—EARNINGS RELEASES: A score that represents the news sentiment of the given story according to the BER classifier, which specializes in news stories about earnings releases. Scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively. This sentiment score is based on RavenPack’s Expert Consensus Methodology.

ANL-CHG—ANALYST RECOMMENDATIONS & CHANGES: A score that represents the recommendation by an analyst in terms of stock upgrades and downgrades. When the mention of a company in a story matches the criteria for ANL-CHG, scores can take values of 0, 50 or 100, indicating a downgrade, neutral or upgrade rating, depending on recommendations issued by the analyst.

MCQ—MULTI CLASSIFIER FOR EQUITIES: A score that represents the news sentiment based on the tone towards the most relevant company mentioned in a story. The score is derived from a combination of values produced by the BMQ, BEE, BCA and ANL-CHG classifiers. An MCQ score is assigned when a company is mentioned in a headline and tagged with a sentiment value by any of these four classifiers. When the mention of a company in a story matches the criteria for MCQ, scores can take values of 0, 50 or 100 indicating negative, neutral or positive sentiment, respectively.

EDITORIAL_NOVELTY: A single news event may often be reported as a chain of linked stories. This number identifies the order of the story in a news chain. Integer scores have a minimum of 1 and no maximum. A score of 1 indicates this is the first take of the story published, whereas a score of 2 indicates this is the second take.

DJ_ACCESSION_NUMBER: This numeric identifier assigned by Dow Jones identifies to which news chain a given story belongs. Stories that are part of the same chain have the same Dow Jones accession number; those that are part of different chains have a distinct accession number.
**RP_STORY_ID**—RAVENPACK UNIQUE STORY IDENTIFIER: An alphanumeric character identifier to uniquely identify each news story analyzed. This value is unique across all records.

**NEWS_TYPE**—TYPE OF NEWS STORY: Classifies the type of news story into one of five categories: HOT-NEWS-FLASH, NEWS-FLASH, FULL-ARTICLE, PRESS-RELEASE and TABULAR-MATERIAL.

*Coding file field descriptions*

**TIMESTAMP_UTC:** The date/time (yyyy-mm-dd hh:mm:ss.sss) at which the news item was received by RavenPack servers in Coordinated Universal Time (UTC).

**RP_STORY_ID**—RAVENPACK UNIQUE STORY IDENTIFIER: An alphanumeric character identifier to uniquely identify each news story analyzed. This value is unique across all records.

**DJ_STORY_ID**—DOW JONES UNIQUE STORY IDENTIFIER: An alphanumeric character identifier provided by Dow Jones to uniquely identify each news story analyzed.

**DJ_INDUSTRY:** Includes metadata tags applied by Dow Jones that identify about 150 industry categories to which the given story relates.

**DJ_GOVERNMENT:** Includes metadata tags applied by Dow Jones that identify to which government bodies, agencies, representatives and personnel the given story relates.

**DJ_NEWS:** Metadata tags applied by Dow Jones that identify to which subject the given story relates.

**DJ_MARKET:** Metadata tags applied by Dow Jones that identify to which market sector or swift currency code the given story relates.

**DJ_REGION:** Metadata tags applied by Dow Jones that identify more than 200 countries, U.S. states, territories and broader regions to which the given story relates.

**DJ_WSJ:** Metadata tags applied by Dow Jones that identify The Wall Street Journal topic code to which the story relates.

**DJ_COMPANIES:** Metadata tags applied by Dow Jones that identify to which of more than 30,000 companies a given story relates.

**DJ_COMPANIES_ISIN:** Includes transformations to ISINs for the metadata tags applied by Dow Jones that identify to which company a given story relates. The ISINs are transformed using an internal database by RavenPack and include any of the companies tracked by RavenPack along with ISINs sent by Dow Jones in the metadata of the story.
(iii) Example of news data in tabular form

**Analytics file: data fields 1–8**

<table>
<thead>
<tr>
<th>TIMESTAMP UTC</th>
<th>COMPANY</th>
<th>ISIN</th>
<th>RP_CO_ID</th>
<th>RELEVANCE</th>
<th>CATEGORY</th>
<th>ESS</th>
<th>ENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-08-05 01:12:47.383</td>
<td>SG/H17</td>
<td>SG1083915098</td>
<td>F051FD</td>
<td>100</td>
<td>analyst-ratings-change-positive</td>
<td>78</td>
<td>100</td>
</tr>
<tr>
<td>2010-08-05 01:13:03.236</td>
<td>US/NLY</td>
<td>US03571O4092</td>
<td>084D10</td>
<td>100</td>
<td>insider-buy</td>
<td>61</td>
<td>75</td>
</tr>
<tr>
<td>2010-08-05 01:13:15.146</td>
<td>SG/H17</td>
<td>SG1083915098</td>
<td>F051FD</td>
<td>100</td>
<td>price-target upgrade</td>
<td>74</td>
<td>100</td>
</tr>
<tr>
<td>2010-08-05 01:14:51.376</td>
<td>JP/5732</td>
<td>JP3866880000</td>
<td>CS4555</td>
<td>100</td>
<td>product-recall</td>
<td>29</td>
<td>100</td>
</tr>
<tr>
<td>2010-08-05 01:16:18.540</td>
<td>US/WG</td>
<td>US5992031084</td>
<td>A8F18</td>
<td>100</td>
<td>revenues</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>2010-08-05 01:16:22.877</td>
<td>US/WG</td>
<td>US5992031084</td>
<td>A8F18</td>
<td>100</td>
<td>earnings-per-share-positive</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>2010-08-05 01:16:26.449</td>
<td>US/WG</td>
<td>US5992031084</td>
<td>A8F18</td>
<td>100</td>
<td>revenues</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>2010-08-05 01:18:29.013</td>
<td>JP/8031</td>
<td>JP3898600001</td>
<td>1AEO41</td>
<td>100</td>
<td>stock-gain</td>
<td>55</td>
<td>100</td>
</tr>
</tbody>
</table>

**Analytics file: data fields 9–23**

<table>
<thead>
<tr>
<th>ENS_KEY</th>
<th>CSS</th>
<th>NIP</th>
<th>PEO</th>
<th>BEE</th>
<th>BMQ</th>
<th>BSM</th>
<th>BCA</th>
<th>BER</th>
<th>ANL_CHG</th>
<th>MCO</th>
<th>EDITORIAL_NOVELTY</th>
<th>DI_ACCESS_NUMBER</th>
<th>RP_STORY_ID</th>
<th>NEWS_TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>G55858..</td>
<td>55</td>
<td>73</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>...22420</td>
<td>655858..</td>
<td>NEWS-FLASH</td>
<td></td>
</tr>
<tr>
<td>89931...</td>
<td>50</td>
<td>43</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>1</td>
<td>...22421</td>
<td>6855B..</td>
<td>FULL-ARTICLE</td>
<td></td>
</tr>
<tr>
<td>337E5...</td>
<td>55</td>
<td>61</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>2</td>
<td>...22420</td>
<td>337E5..</td>
<td>NEWS-FLASH</td>
<td></td>
</tr>
<tr>
<td>6672C...</td>
<td>50</td>
<td>26</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>1</td>
<td>...22427</td>
<td>6672C..</td>
<td>NEWS-FLASH</td>
<td></td>
</tr>
<tr>
<td>0BB9A...</td>
<td>55</td>
<td>43</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>1</td>
<td>...22435</td>
<td>0BB9A..</td>
<td>PRESS-RELEASE</td>
<td></td>
</tr>
<tr>
<td>50254...</td>
<td>50</td>
<td>52</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>3</td>
<td>...22435</td>
<td>60254..</td>
<td>NEWS-FLASH</td>
<td></td>
</tr>
<tr>
<td>0BB9A...</td>
<td>52</td>
<td>56</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>100</td>
<td>4</td>
<td>...22435</td>
<td>4657D..</td>
<td>NEWS-FLASH</td>
<td></td>
</tr>
<tr>
<td>G8128...</td>
<td>62</td>
<td>35</td>
<td>100</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>100</td>
<td>1</td>
<td>...22451</td>
<td>G8128..</td>
<td>FULL-ARTICLE</td>
<td></td>
</tr>
</tbody>
</table>

**Coding file: data fields 1–7**

<table>
<thead>
<tr>
<th>TIMESTAMP UTC</th>
<th>RP_STORY_ID</th>
<th>DJ_PROJECT_ID</th>
<th>DJ_INDUSTRY</th>
<th>DJ_GOVERNMENT</th>
<th>DJ_NEWS</th>
<th>DJ_MARKET</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-08-05 01:12:47.383</td>
<td>655858..</td>
<td>...22420</td>
<td>/FAC.,/XDI.,/GL.,/</td>
<td>N/AER,N/ANL/N/..</td>
<td>M/DIUM/MMR,..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:13:03.236</td>
<td>6855B..</td>
<td>...22421</td>
<td>/BEA,/R11U,..</td>
<td>G/CNG,G/AUS,..</td>
<td>N/CNW,N/DIN,N/..</td>
<td>M/FCL,M/NND,M/..</td>
</tr>
<tr>
<td>2010-08-05 01:13:15.146</td>
<td>337E5..</td>
<td>...22420</td>
<td>/FAC.,/XDI.,/GL.,/</td>
<td>N/AER,N/ANL/N,..</td>
<td>M/DIUM,NNND,..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:14:51.376</td>
<td>6672C..</td>
<td>...22427</td>
<td>/CSE,/ENT,/..</td>
<td>N/ADR,N/AER,N/..</td>
<td>M/MMR,M/MCY,..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:16:18.540</td>
<td>0BB9A..</td>
<td>...22435</td>
<td>/ECO,VOI,..</td>
<td>N/CNW,N/DIF,N/..</td>
<td>M/ENE,M/MMR,M/..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:16:22.877</td>
<td>60254..</td>
<td>...22435</td>
<td>/ECO,VOI,..</td>
<td>N/CAC,N/CNW,N,..</td>
<td>M/ENE,M/MMR,M/..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:16:26.449</td>
<td>4657D..</td>
<td>...22438</td>
<td>/ECO,VOI,..</td>
<td>N/CAC,N/CNW,N,..</td>
<td>M/ENE,M/MMR,M/..</td>
<td></td>
</tr>
<tr>
<td>2010-08-05 01:18:29.013</td>
<td>G8128..</td>
<td>...22451</td>
<td>/DIB,USCV,..</td>
<td>N/ADR,N/AER,N/..</td>
<td>M/DIUM/MMND,..</td>
<td></td>
</tr>
</tbody>
</table>

**Coding file: data fields 8–11**

<table>
<thead>
<tr>
<th>DJ_REGION</th>
<th>DJ_WEIS</th>
<th>DJ_COMPANIES</th>
<th>DJ_COMPANIES_ISIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>R/FE,R/PO,E,R/..</td>
<td>H17.5G</td>
<td>SG1083915098</td>
<td></td>
</tr>
<tr>
<td>R/NL,R/NME,R/U,..</td>
<td>NLY</td>
<td>US03571O4092</td>
<td></td>
</tr>
<tr>
<td>R/FE,R/PO,E,R/..</td>
<td>H17.5G</td>
<td>SG1083915098</td>
<td></td>
</tr>
<tr>
<td>R/ASL,R/CH,R/..</td>
<td>J/SFR</td>
<td>6752.TO</td>
<td>J/P386688000000</td>
</tr>
<tr>
<td>R/NME,R/RT/R,U,..</td>
<td>KG</td>
<td>US9920310084</td>
<td></td>
</tr>
<tr>
<td>R/NME,R/RT/R,U,..</td>
<td>J/MKU,0/TVV,..</td>
<td>WIG</td>
<td>US5992031084</td>
</tr>
<tr>
<td>R/NME,R/RT/R,U,..</td>
<td>J/CMU,L/MIM,..</td>
<td>WIG</td>
<td>US5992031084</td>
</tr>
<tr>
<td>R/ASR,R/FE,R/..</td>
<td>B0831.TO,MITSY</td>
<td>J/P3898600001</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.10. Data layout from RavenPack News Analytics—Dow Jones Edition.
1.B REFERENCES


Das S. (this volume) “News analytics: Framework, techniques, and metrics” (see Chapter 2).


diBartolomeo D. (this volume) “Using news as a state variable in assessment of financial market risk” (see Chapter 10).


Dzielinski M.; Rieger M.O.; Talpsepp T. (this volume) “Volatility, asymmetry, news, and private investors” (see Chapter 11).


Kalev P.S.; Duong H.N. (this volume) “Firm-specific news arrival and the volatility of intraday stock index and futures returns” (see Chapter 12).


Kittrell J. (this volume) “Sentiment reversals as buy signals” (see Chapter 9).


Leinweber D.; Sisk J. (this volume) “Relating news analytics to stock returns” (see Chapter 6).


Lo A.; Healy A. (this volume) “Managing real-time risks and returns: The Thomson Reuters NewsScope Event Indices” (see Chapter 3).


