|  |
| --- |
|  |
| News Analytic Applications in the Financial Industry |
|  |
| Literature Review |

|  |  |
| --- | --- |
| Name: | Yan Dawei |
| Matric No: | 1001433J |
| Major Project Group Name: | SOLY |
| Project Title: | nTrader |

Contents

[1. Introduction 1](#_Toc342471137)

[2. Financial Applications 1](#_Toc342471138)

[2.1 Information Flow and computational architecture 2](#_Toc342471139)

[2.1.1 Input 2](#_Toc342471140)

[2.1.2 Transformation 3](#_Toc342471141)

[2.1.2.1 Conversion of qualitative text 5](#_Toc342471142)

[2.1.2.2 Classifiers 6](#_Toc342471143)

[2.2 Trading and fund management 8](#_Toc342471144)

[2.2.1 Stock picking and ranking 8](#_Toc342471145)

[2.3 Monitoring Risk and Risk control 9](#_Toc342471146)

[3. Summary 9](#_Toc342471147)

[Bibliography 11](#_Toc342471148)

[Appendix 0 (Figures) 12](#_Toc342471149)

[Appendix 1 Overview of the model components 13](#_Toc342471150)

[Appendix 2 Construction of the Lexicon 14](#_Toc342471151)

[Appendix 3 Discriminant Values 15](#_Toc342471152)

Table of Figures

[Figure 1: Information flow and computational architecture: Taken from Mitra 2010 12](file:///C:\Users\Yan\Desktop\News%20Analytics%20in%20the%20Financial%20Industry.docx#_Toc342468104)

[Figure 2: Sentiment Extraction algorithm and system design 12](#_Toc342468105)

Abstract

Purpose of this Literature Review is to explore the subject news analytics and how it is being utilized in the financial industry so that nTrader can explore alternatives when building its own news analytical applications. A brief introduction of the subject will be given followed by its applications & strategies adopted by the financial industry.

# Introduction

“News Analytics refers to the measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories” – *Wikipedia.*

News which is considered to be the main source of current events in the global economy is monitored by many to make decisions and identify opportunities. A company which wants to undergo expansion can seek out the latest developments in the various emerging countries through the news so as to spot potential markets to tap into. An individual who wishes to invest in the financial markets can find out which economy in the world is doing relatively better through the news. And the list of examples goes on.

With the birth of the internet, the accessibility and supply of global news contents increased dramatically. With that much more information, individuals can now assess their subject of interests more comprehensively. However time needed to digests and analyse these information varies from individual to individual as the level of comprehension skill of each individual is different. Furthermore, analysing text contents are subjective as different people may have different views on a particular news article. Thus, in order to stay on top of the game, big financial corporations had to figure out a way to analyse huge amounts of news contents simultaneously, objectively and at rapid speed. Leveraging on the breakthrough of natural language processing, News Analytics software that are being supplied by prominent household names such as DOW JONES and Thomson Reuters are quickly being adopted by hedge funds, mutual funds and trading firms to capture profitable opportunities in the financial markets. In the below sections we will be exploring the various usages of news analytic in the financial industry in detail.

# Financial Applications

In Finance, within the investment management process, the role of News Analytics is to process news by leveraging on technology and algorithms. This technology allows investors or traders to surpass the limits of the investment process by automating human thinking and reasoning.

In order to predict or anticipate the possible direction of asset returns and the level of uncertainty, traders, speculators or private investors will have to read and interpret recent economic and financial news to gain a deeper understanding of the current economic climate. 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. What news analytics does is to automate or semi-automate this whole process. By automating this judgement process, the human decision maker can act on a larger, hence more diversified, collection of assets.

In increasingly competitive markets, traders and investors need to select and analyse the relevant news, from the massive quantities available to them, in order to make smart and timely decisions. However, one human or even a group human’s ability to absorb and comprehend news is limited. With news analytics technology, these humans are able to extract, aggregate and categorise large volumes of news effectively. Application areas in the finance industry will include but not limited to:

* **Trading**
* **Fund Management**
* **Monitoring risk and risk control**

We will be exploring the above application areas briefly in later sections.

The benefit from automated news analysis is limitless, from reducing the time taken for traders to react to breaking stories, to automated filtering, monitoring and aggregation of news, effectively reducing the burden of routine monitoring and allowing them to better target their reading and research. To sum it all up, news analytics technology can be geared towards human decision support as well as be used to create automated quantitative strategies. The combination of news data with historic market data makes models more proactive and less reactive (Mitra, 2010).

## Information Flow and computational architecture

Figure 1 in Appendix 0 depicts a common information flow and the corresponding Information system architecture of a News Analytics incorporated financial system application (Mitra, 2010). In the below sub-sections we will be exploring what are the various types of news data and how can news data be transformed into more meaningful investment data.

### Input

There are two main types of financial news:

* **Regular synchronous announcements(scheduled or expected news)**
* **Event driven asynchronous announcements(unscheduled or unexpected news)**

Regular synchronous announcements refer to the source data that reports research before they write news articles. This data comes from primary information sources such as the reports and filings from the various global securities commissions, court documents and government agencies. It also includes scheduled announcements such as macro-economic news, industry statistics, company earnings reports and other corporate news. This type of news is often very structured, having a well-defined numerical and textual content. In the largest and most liquid markets, such as foreign exchange, government debt and futures markets, firms often execute large and rapid trading strategies using economic indicators as input. As these news events are normally very well structured and documented, thorough back testing of strategies is feasible. Because of the precise scheduling of this type of news, market participants are normally very well prepared to deal with them therefore speed and accuracy are the major determinants of success when deploying such strategies. Earnings report is also widely anticipated news and a key input for trading strategies. In Singapore, SGX is the primary provider for earnings reports of all listed companies.

Event driven asynchronous announcements are news that surface without a fix schedule. They often surface as textual, unstructured and qualitative data. Examples of such are Main stream news (newspapers, radio or television), rumours (blogs and websites that broadcast “news”) and social media. Unlike the former, this type of news data contains information about the effect of an event and the possible causes of an event. In order for it to be applied in trading systems and quant models, it needs to be transformed into quantitative inputs. One good technique is to covert the data into a quantitative time series[[1]](#footnote-1). Many aspects of news can be converted, some of which will include (Mitra, 2010):

* **A binary series where the occurrence of a particular event or the publication of a news article about a particular topic is indicated by 1 and the absence of the event can be indicated by 0.**
* **Measuring news flow (volume of news)**
* **Determine scores (measures) based on the language sentiment of text**
* **Determines scores based on the market’s response to particular language**

Analyses done on these quantitative inputs are somewhat similar to technical analysis done on price charts. The various ways that the raw news data will be analysed and converted will be further discussed in the next phase, Transformation.

### Transformation

At this stage, Apart from conversion to quantitative inputs, extracted news data will also be pre-analysed and considerations to look into will include (Mitra, 2010):

* **Differences in the availability of news data for different companies**

Much research done has found out that those larger companies, as defined by larger market capital, has higher news coverage (Moniz, Brar and Davis 2009, Cahan,Jussa and Luo 2009). With this intelligence, the trader can choose to devote more time for analysis on these larger capped companies.

* **Classifications of news items**

To increase the efficiency and ease of executing news trading strategies and back testing, news articles should be tagged and classified into different genres. Fortunately, many news providers have already incorporated this feature into their news which aids executing event based trading strategies. Tagging news articles in different event tags is important as this allows us to distinguish what type of news is relevant to our quantitative or financial model. Further analysis can be conducted on these different types of events to conclude which event has a greater reaction from the markets. Other than event types, a group of articles can also be grouped accordingly to a certain company tag.

* **Sieve out relevant and current news**

This is a very important as we will not want an algorithm specifically designed for a particular industry or company to be processing on a piece of irrelevant data pertaining to another industry. That kind of mistake may just costs millions of dollars in losses when the wrong trading signal is generated. For an event based trading strategy, we would also want to filter out news data that are not an event in nature or have events that have similar degree of impact or sentiment. Further filtering can be done to minimise the number of duplicate stories. This can be done by measuring the uniqueness of each news article by comparing to previous news articles, effectively revealing how many similar articles there are for a particular company. Based on studies, markets react more strongly when “new” or fresh news is released, therefore it is important as well to identify such news, e.g. new product releases (Moniz, Brar and Davis 2009).

* **Variations in volume of news flow**

This refers to the periods of unexpected news flow levels or periods of variation due to seasonality. Identifying these periods will render further investigation and findings or reasons for such periods of news flow patterns may aid us in developing seasonal or news flow oriented trading strategies. According to Hafez’s research, a RavenPack[[2]](#footnote-2) personnel, he discovered that larger volumes of news flow arrive just before the opening of the European, US and Asian trading sessions while little news flow takes place on weekends. In the week, the peak of news flow occurs on Wednesday and Thursday, while the trough falls on Friday.

* **Time of the day when news is released**

According to the research article ~ *What type of events provide the strongest evidence that the stock market is affected by company specific news ~* , a greater likelihood of events that lead to rising volatility at the start of the day. This fact proves that the time of the day when news is released is relevant in understanding the connection between market variables and news, rendering the need for further investigation.

* **Market conditions that can influence the types of news reported**

According to Boyd, Hu and Jangannathan (2005), interest rate information dominates in expansionary periods. In contrast information about future corporate dividends dominates when the markets are contracting. Such intelligence is useful when considering the impact of news on financial markets at different stages of the business cycle. A study done by Bestelmeyer and Hess suggested that during recessions, good news from employment report is good news for the stock market, however during expansions; good news from the labour market is bad news for the stock market.

* **Informational content of news**

This refers to the sentiment of the news, which is expected to have a large influence on how markets react to news (Bestelmeter and Hess 2010, Tetlock 2007). According to Tetlock, stock returns react more strongly to pessimistic news than optimistic ones. Converting qualitative text into qualitative metrics (sentiments) will be discussed in the next section.

* **Relationship of different news stories to each other**

Since we know that news are able to move the financial markets, the dependence and independence of types or classifications of news articles will be of a great benefit. One such example would be finding out leading and lagging events. A leading event of a company can be a news article about profit warning; possible subsequent lagging news that can follow up can be regarding bankruptcy or resignation. With this intelligence, predictive capabilities will be possible. Therefore finding out how different news articles are related and in what degree demands serious investigations.

The above examples of business intelligence on news data will be very beneficial for the development of quantitative models, fund management strategies as well as risk management functions.

### Conversion of qualitative text

As mentioned in the sections above, qualitative text has to be converted into quantitative inputs in order to be meaningful data that can be used by quantitative models. As discussed, the informational content or the sentiment of news is the most important aspect of news analysis in the area of financial application. In order to distinguish whether an article’s content is positive or negative and by how much, we have to be able to find a way to assign a quantified sentiment score or index to it. One industry standard technique to achieve this is to capture the emotive content of the article through the use of sophisticated machine learning and natural language algorithms. This allows us to determine sentiment scores over times as news arrives. Once a sentiment index is constructed, to use it effectively, we have to find evidence of its relationship with relevant asset returns, trading volumes or volatility. We will explore the conversion methodology in detail below.

An overview of the whole methodology flow chart can be found in the appendix 0 figure 2. News contents are first downloaded from the internet through a “Web-scraper” program and fed to algorithms to be classified into either one of the following: bullish (optimistic), bearish (pessimistic), and neutral (spam or content that are neither of the former two). There are four algorithms in total, each with different theoretical foundations to classify each message. A training corpus will be established to test the algorithm’s performance on a small subset of pre-classified news data. This training corpus will be kept deliberately small to avoid over fitting[[3]](#footnote-3). The algorithms then “learn” sentiment classification rules from the pre-classified data set, and will apply these rules out-of-sample. A simple majority across the four algorithms is required before a news content is finally classified, or else it is discarded. This voting approach will ensure that results are more accurate for determining sentiments.

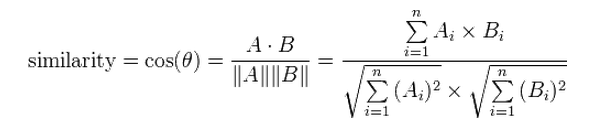
Three complementary databases support the classification algorithms (Das and Chen 2007):

* An English “dictionary”, which provides base language data to determine the nature of the word. I.e., noun, adjective, adverb, etc.
* A “lexicon” which is hand-picked collection of finance words (such as bull, bear, uptick, value, buy, pressure, etc.). These words form the variables for statistical inference undertaken by the algorithms. For example, when we count positive and negative words in an article, we will use only words that appear in the lexicon, where they have been pre-tagged for sign.
* A “grammar” or the pre-classified training corpus. It forms the base set of news data for further use in classification algorithms. These pre-classified news data provide the in-sample statistical information for use on the out-of-sample messages.

Appendix 1 will describe these three databases explicitly. The five algorithms presented in this literature review will make use of these three algorithms to arrive at the three-way classification of news article.

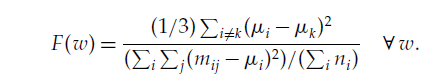
### Classifiers

1. **Naïve Classifier** – This algorithm is based on a word count of positive and negative implication words. Each word in a news article is checked against the lexicon, and assigned a value (-1, 0, +1) based on the default value (sell, null, buy) in the lexicon. After which, a net word count of all lexicon matched words is taken, and if this value is >1, we sign the article as a buy. If this value is <1, it is taken as a sell. All others are treated as neutral (Das & Chen, 2007).
2. **Vector Distance Classifier** – This algorithm tries to match the news article to its closest match within the pre-classified training data. Take for example there are 300 words in the lexicon, each word will be assigned a dimension in vector space, the lexicon will then be like a 300-dimensional unit hypercube. Every article can be thought as a word vector in this space. The elements in the vector takes values in the set {0, 1, 2, ….} depending on how many times a word appears in the article. The same will be applied to the training corpus, where the pre-classified data are all converted into vectors. The article will then be classified by comparison the group of pre-trained vectors .



As can be seen from the equation above, each article (*A*) will be assigned the classification of the grammar rule (*B*) with which it has the smallest angle or most similarity with. A variation on the theme could be to use sets of top-*n* closest vectors (Das & Chen, 2007).

1. **Discriminant-Based Classifier –** This algorithm is an advanced form of the naïve classifier whereby lexical words are used to determine the sentiment of the article. However, instead of treating words with equal importance, this classifier gives a different weighting to each words depending on how indicative the words are. Using the training corpus, we compute a measure of the discriminating ability of each word in the lexicon. After which, the simple word count in the naïve classifier is replaced by weighted word count.

****

The equation goes like this: Let the mean score (average number of times word *w* appears in a text message of category *i*) of each term for each category =*ui*, where *i* indexes category (neutral, positive, negative). Let text messages be indexed by *j*. The number of times word *w* appears in a message *j* of category *i* is denoted *mij*. Let *ni* be the number of times word *w* appears in category *i*.

The equation above assigns a score to F(*w)* to each word in the lexicon, which is the ratio of the mean across-class squared variation (class *i* (positive)vs class *k* (negative)) to the average of within-class squared variation (class *i* (positive)). Let’s use an example to get some intuition. If the word “crisis” appears exactly once in a text that is pessimistic and never appears in text that is optimistic, then the within-class variation is zero, and the across-class variation is positive. In such a case, where the denominator of the equation above is zero, the word “crisis” is an infinitely powerful discriminant. It will be given a very large weight.

Each word in the article will be checked against the lexicon, and assigned a signed value (-*v, 0, +v*), based on the sign (sell =-1, null =0, buy =+1) in the lexicon multiplied by the discriminant value *v*=F(*w*). After this assignment, the net word count of all lexicon-matched words is taken, and if this value is greater than 0.01, the article is a “buy”. If the value is less than -0.01, the article is taken as a “sell”. All other articles are treated as neutral again (Das & Chen, 2007).

1. **Adjective-Adverb Phrase Classifier** – The assumption in this algorithm is that adjectives and adverbs emphasize sentiment and require greater weight in the classification process. The goal of this algorithm is to only focus on words in specially chosen phrases containing adjectives and adverbs. The algorithm will make use of the dictionary to detect noun phrases containing adjectives or adverbs. Once this is detected, we form a “triplet”, which consists of the adjective or adverb and the two words immediately following or preceding it in the message. This triplet usually contains meaningful interpretive information because it contains the adjective or adverb, both of which are parts of speech that add emphasis to the phrase in which they are embedded. After identifying these phrases, the lexicon will then be used to determine whether these connote positive or negative sentiment. A net count in these phrases will determine whether it is positive, negative or neutral (Das & Chen, 2007).

## Trading and fund management

What most traders and quantitative fund managers does is identifying and exploiting asset mispricing, before they correct, in order to generate alpha. By making use of news analytics, they can use quantified news data to rank stocks and identify which stocks are relatively attractive or unattractive. They may then buy or sell the highest or lowest ranking stocks, thereby rebalancing a portfolio composed of desired weights on the selected stocks. Similarly the news data can also be used to identify trading signals for particular stocks. Factors models[[4]](#footnote-4) are an alternative that analysts may use to process new sources of news data. Analysts may also use news data to identify and exploit behavioural biases in investor behaviour arising due to the market and analysts’ untimely or incorrect reaction to new information. This can be caused by delayed information dissemination or due to investors’ inattentiveness and limited capability to process all relevant information instantly (Mitra, 2010).

### Stock picking and ranking

Li (2006) uses a simple ranking procedure to identify stocks with bullish or bearish sentiment. By examining from 10-K Securities and Exchange Commission (SEC) filings for non-financial firms between 1994-2005, he managed to create 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. He then devised a strategy to go long on stocks with low risk sentiment measure and short stocks with high risk sentiment measure which he found to produce a reasonable level of returns.

## Monitoring Risk and Risk control

For effective financial risk control, companies need to identify, understand and quantify potential adverse outcomes, their related probabilities and the severity of their impacts. Traditional approaches of using historic asset price data fails to account for developments in the market environment, investor sentiment and knowledge. Since market conditions are likely to vary from historic observations, it is important to incorporate measures or observations of the market conditions within the estimation of future portfolio return distributions in order to capture the true level of risk.

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

# Summary

The development of news analytics and its applications to finance through sentiment analysis is gaining progressive popularity within the investment community. In this paper I have covered briefly how news analytics is being carried out in the financial industry as well as their applications to trading, fund management and risk control:

* Risk management – Use of news data within risk forecasting to enable dynamic risk management strategies that are forward looking and are based on changing market environments. This risk analysis applied using news data can also help investors understand event risk and how different kinds of events can impact their portfolio risk profile.
* Stock screening tool – Sentiment data may be used to predict the directional movement of future returns. Stocks that of high positive sentiment may be held long while those of low positive sentiment may be held short.
* Trader decision support – News data signals can be used to confirm traders existing analysis or it may cause them to reconsider their analysis.

# Bibliography

A-Team Group. (October, 2009). *A-Team Group.* Retrieved 30 November, 2012, from A-Team Group Website: https://docs.google.com/viewer?a=v&q=cache:Ky3qAuf7FR4J:www.hedgehogs.net/mod/file/download.php%3Ffile\_guid%3D1080386+&hl=en&gl=sg&pid=bl&srcid=ADGEESiR3dqWsO-ZBF1\_P8DXj6-P9CRsqiZv6UU0CK32YwcenkhbKXpFJExL9TxM0Vj2Yf2RqbHQPc7VaGGfX3nSgLfE6iOMNHSUcCoyiRPwJGz

Georg Bestelmeyer, D. H. (January, 2010). Retrieved 2 December, 2012, from http://www.frankfurt-school.de/clicnetclm/fileDownload.do?goid=000000318381AB4

Leela Mitra, G. M. (17 June, 2007). *OptiRisk Systems.* Retrieved 28 November, 2012, from OptiRisk Systems web site: http://optirisk-systems.com/papers/Opt0014.pdf

Leinweber, D. (2009). Retrieved 1 Decemer, 2012, from https://www.scss.tcd.ie/Khurshid.Ahmad/Research/High\_Frequency\_Trading/Nerds%20on%20Wall%20Street.pdf

Li, F. (21 april, 2006). *Social Science Research Network.* Retrieved 29 november, 2012, from Social Science Research Network web site: http://papers.ssrn.com/sol3/papers.cfm?abstract\_id=898181

Marc-André Mittermayer, G. F. (2007). Retrieved 2 December, 2012, from http://www.iadis.net/dl/final\_uploads/200708C025.pdf

PAUL C. TETLOCK, M. S.-T. (n.d.). Retrieved 1 December, 2012, from http://www.haas.berkeley.edu/groups/finance/TSM\_More\_Than\_Words\_02\_07.pdf

Sanjiv R.Das, M. Y. (20 July, 2007). *Sanjiv Ranjan Das.* Retrieved 28 November, 2012, from Sanjiv Ranjan Das: http://algo.scu.edu/~sanjivdas/chat\_FINAL.pdf

TETLOCK, P. C. (n.d.). Retrieved 2 December, 2012, from http://www.afajof.org/afa/forthcoming/2471.pdf

# Appendix 0 (Figures)

Mainstream news

Pre-News

Web 2.0

Social Media

(Numeric) financial market data

Updated beliefs, Ex-ante view of market environment

Quant Models

1. Return Predictions
2. Fund Management/ Trading Decisions
3. Volatility estimates and risk control

Consolidated Analysis

Sentiment scores

News Analysis

Figure 1: Information flow and computational architecture: Taken from Mitra 2010

Consolidated Datamart

Database

Helper programs:

-Message parsing

-Dictionary Handler

Web-Scraper Program

Preprocessing:

Clean up HTML, expand abbreviations, tag negations

Discriminant Classifier

Naïve Classifier

Vector Distance Classifier

Adjective/Adverb Classifier

Classified messages

Figure 2: Sentiment Extraction algorithm and system design

# Appendix 1 Overview of the model components

A1.1 **Dictionary**

The data here contains supplementary information on the English Language. This dictionary will assist in exploiting parts-of-speech usage in articles; detecting adjectives and adverbs for the classifier algorithms. There are a few such databases that are available for free to the public; examples would be CUVOALD[[5]](#footnote-5) and WordNet[[6]](#footnote-6). Such databases in general are called lexical databases[[7]](#footnote-7).

A1.2 **Lexicon**

The lexicon is designed using domain knowledge and statistical methods, containing a group “discriminant” words. A discriminant function is used to statistically detect which words in the training corpus are good candidates for classifier usage (refer to section regarding algorithms). To sum it up, the lexicon is essentially a collection of words relevant to the classification problem, which will be used by the classifier algorithms to discriminate buy sentiments from sell sentiments. Our goal here is to populate the lexicon with words of high discriminant value. Over time, more words may be added to the lexicon, which improves in this evolutionary manner. More details on the lexicon are presented in Appendix 2.

A1.3 **Grammar**

The grammar is can be defined as a set of functions or rules applied in conjunction with the lexicon to extract sentiment from text. Correspondences between words sets, language features, and classification types comprise the grammar. In this context, the training data is the grammar. These pre-classified in-sample data can be thought of as a set of rules that govern the classification of out-of-sample data. Take for example, we wish to analyse article *A*, using some metric, we can compare the relationship of article *A* to a set of other pre-classified articles *B* and find the one that is its closest match. The properties of articles *A* will then be equated to those of the proxy. The set of pre-classified articles *B* is in this case, the grammar, and the rule that finds the proxy article or a proxy set of articles is codified in a classification algorithm. The classification algorithm implements a rule that seeks out closest articles in the grammar, using the words in the lexicon as variables. Some of the algorithms use only the grammar, or the lexicon, and some use both (Das and Chen 2007).

A1.4 **Message Pre-Processing**

Before applying the lexicon-grammar based algorithms, each news article is pre-processed to enable cleaner interpretation. HTML tags will be removed as they may be concatenated to lexical items of interest. Abbreviations will be expanded to their full form. For example, the word “isn’t” is replaced with “is not”. Negations words will also be handled appropriately. Negation words normally cause the meaning of the sentence to be the opposite of that without the negation. For example, the sentence “it is not a bearish market” actually means the opposite of a bearish market. Words such as “not”, “never”, “no”, etc., serve to reverse meaning. These words are detected and the rest of the words in the sentence after are marked, so as to reverse inference. These three parsers deliver a clean set of messages for classification.

# Appendix 2 Construction of the Lexicon

The features of the lexicon are as follows:

* These words should be hand-selected based on the reading of a significant amount of sample data.
* The lexicon is user-specified, allowing the methodology to be tailored to individual preference. For example, if the user is only interested in articles that relate to stocks, a lexicon containing mostly stock-related words may be designed. (The grammar, i.e. the training set would also be correspondingly tagged.)
* For each word in the lexicon, we tag it with a “base” value, i.e. the category in which it usually appears. For example, the word “sell” would be naturally likely to appear in messages of type SELL, and we tag “sell” with base value 1. If the word is of BUY type, we tag it with value 3, and NULL words are tagged 0. Every time a new word is added to the lexicon, the user is required to make a judgement on the base type.
* Each word is also “expanded,”i.e. appears in the lexicon in all its forms, so that across forms, the word is treated as one word. This process is similar to stemming[[8]](#footnote-8) words, except that we exhaustively enumerate all forms of the word rather than stem them.
* Each word is also entered with its “negation” counterpart, i.e. the sense in which the word would appear if it was negated. Negation is detected during pre-processing and is used to flag portions of sentences that would be reversed in meaning. An example of a lexical entry along with its base value, expansion and negation is provided below:
* 3 favourable favourite favourites favouring favoured
* 1 favourable\_n favourite\_n favourites\_n favouring\_n favoured\_n

All forms of the word appear in the same line of the lexicon. As can be seen, a tag is attached to each negated word in the second line above. The default classification value (the “base” value) is specified at the beginning of the line for each lexical item (i.e. a 0, 1, or 3).

Based on the training corpus, we can compute the discriminant value of each item in the lexicon. This value describes the power of the lexical item in differentiating sentiment types. For example, the word “sell” is likely to be a strong discriminator, since it would be suggestive of negative sentiment. The goal is to populate the lexicon with words that are good discriminators. Appendix 3 will provide a brief description of discriminant values.

# Appendix 3 Discriminant Values

Example values for some words from the discriminant function are shown here. The last three words appear with their negation tags (Das and Chen 2007).

SAMPLE DISCRIMINANT VALUES

bad 0.040507943664639216

hot 0.016124148231134897

hype 0.008943543938332603

improve 0.012395140059803732

joke 0.02689751948279659

jump 0.010691670826157351

killing 0.010691670826157329

killed 0.016037506239236058

lead 0.003745650480005731

leader 0.0031710056164216908

like 0.003745470397428718

long 0.01625037430824596

lose 0.12114219092843743

loss 0.007681269362162742

money0.15378504322023162

oversell 0.0

overvalue 0.016037506239236197

own 0.0030845538644182426

gold\_ \_n 0.0

good\_ \_n 0.04846852990132937

grow\_ \_n 0.016037506239236058

The following can be derived from the above sample; the word “lose” understandably has a high discriminant value. The word “oversell” is not used at all. One of the higher values comes from the negated word “good\_\_n” which means that there is plenty of negation in the language used in news articles. Compare this with its antonym “bad”, which actually has a lower discriminant value! The word “joke” is a good discriminator, which is somewhat surprising. The highest valued discriminant is the word “money”.

1. Time series analysis relies on historical data and attempts to project historical patterns into the future [↑](#footnote-ref-1)
2. RavenPack is a leading provider of real-time news analysis services. [↑](#footnote-ref-2)
3. Over fitting decreases efficiency of the model as it tries to memorize training data. Such models will lack in predictive power when presented new or unseen data. [↑](#footnote-ref-3)
4. 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 [↑](#footnote-ref-4)
5. Computer Usable Version of the Oxford Advanced Learner’s Dictionary [↑](#footnote-ref-5)
6. A lexical database which groups English words into sets of synonyms called synsets, provides short general definitions, and records the various semantic relations between these synonym sets [↑](#footnote-ref-6)
7. A database that contains lexical category and synonyms of words, as well as semantic relations between different words or sets of words [↑](#footnote-ref-7)
8. **stemming** is the process for reducing inflected (or sometimes derived) words to their stem, base or root form [↑](#footnote-ref-8)