THOMSON REUTERS NEWS ANALYTICS

SENTIMENT ANALYSIS, RELEVANCE, NOVELTY

WHITE PAPER

Document Version 1.1
Date of issue: February 2013

THOMSON REUTERS™

CONTENTS

CHAPTER 1 INTRODUCTION

News moves markets. Up-to-date knowledge and interpretation of the events that shape the financial landscape provide crucial tools for human traders to generate value. Until now, algorithmic trading has been unable to capitalize on news feeds as a source of tradable information, but with the advent of *automatic sentiment analysis*, the situation has changed.

Analysis of the author sentiment of documents is an established issue-management tool that is employed within a wide range of industries and corporate functions including: media evaluation, market research, brand and reputation management, and political analysis. Traditionally a manual process, recent advances in technology in the field of *computational linguistics* has meant that automatic systems for measuring the author sentiment of documents, news articles and web pages using computers are now possible. The use of technology has obvious speed benefits over manual analysis and allows decision makers to quickly gauge how on their organization is viewed in the wider world and make better decisions.

In the same way, *automatic sentiment analysis* technology can be used to aid algorithmic trading systems make buy/sell decisions in the financial markets by providing a numeric representation of news. With the benefits of speed, accuracy and consistency, the information contained in news can now be consumed by algorithmic trading systems, providing a signal that is orthogonal to market derived signals.

Thomson Reuters News Analytics (TRNA) provides state-of-the-art automated news analysis indicators for algorithmic trading systems. There are a number of indicators: The *TRNA's News Analytics engine* with its *sentiment* and *relevance scoring mechanisms* provides a means of measuring the polarity and magnitude of a news signal; news volatility is indicated by the *novelty scores*; company news volume information can be easily be derived from the TRNA output.

This paper describes the technology behind the News Analytics engine. It starts by describing what we mean by Sentiment Analysis, it then discusses possible approaches to sentiment analysis, before describing Thomson Reuters's solution. Specific design issues are discussed in more detail to explicitly answer common questions raised about the sentiment engine. We then go on to discuss the relevance and novelty calculations.

CHAPTER 2 SENTIMENT ANALYSIS

This section describes what we mean by sentiment analysis.

2.1 WHAT IS SENTIMENT?

We start by defining what it is that we are measuring with a sentiment analysis tool. In Thomson Reuters's opinion the most useful and practical measure that can be extracted from text is that of the sentiment expressed by the author about the subject matter being discussed. This avoids the problem of trying to identify any interpretation that may be put on the facts and opinions within the text by the reader. There are many readers and they may all have different opinions on what the implications of a piece of text are, but they'd generally agree on the author's interpretation. The further implications of the sentiment analysis scores is of course open to interpretation by consumers of the scores, but it is clear that extracting the author sentiment provides a consistent view.

To illustrate what we mean, we'll use an exaggerated example. Consider the following sentences:

1. 'An explosion occurred in Iraq today killing 20 people'
2. 'A horrific explosion occurred in Iraq today murdering 20 people'

The sentiment in the first would be regarded as neutral, even though the event itself might be seen by many people to be objectively negative. However, the second sentence would be negative since the author is using sentiment bearing words to describe the event.

In this way News Analytics extracts author sentiment of the company's performance, that is, whether the author thinks the company is doing well or not. It does not try to gauge market sentiment.

2.2 DOCUMENT AND ENTITY LEVELS

In designing the sentiment engine, Thomson Reuters realised that it was important to specify that sentiment has a subject or target: Sentiment is expressed about something or with regard to something.

Generally, it is possible to perform sentiment analysis over an entire document, termed *document-level* sentiment analysis. If a particular piece of text is about a single subject or event, then measuring the sentiment of the whole text gives a measure of sentiment with respect to that subject or event. For example, in the case of news articles, they tend to describe a particular story. It is valid to say that a measurement of the sentiment of a news article corresponds to the author's opinion to the news story contained within.

However, we can go further than that: In any particular text several opinions may be expressed about multiple subjects. For example, an article about the performance of two companies may be positive about one and negative about the other. For maximally useful sentiment signals, it is important to be able to extract multiple sentiments values for different subjects within a text. Extracting sentiment with respect to entities talked about within the text is termed *entity-level* sentiment analysis.

The News Analytics engine currently performs entity-level sentiment analysis with respect to companies and energy/material topics (e.g. Crude Oil) that are mentioned by the journalist in the metadata that accompanies the story. In this way, the signal is a measurement of sentiment expressed directly towards a tradable entity.

CHAPTER 3 OTHER APPROACHES

There are a number of computational linguistic techniques that have been employed to investigate sentiment analysis in academic literature. Here, we provide an overview of the approaches and their advantages and disadvantages

3.1 LEXICAL ANALYSIS

The simplest approach to Sentiment Analysis involves creating lists, or lexicons, of individually scored words and scoring a piece of text according to the relative frequencies of positive and negative words found in the text.

For example, words such as "good" and "fantastic" may be ascribed positive sentiment, whilst "bad" and "horrible" may be ascribed negative sentiment.

Text is scored by effectively counting the positive and negative words, and calculating the difference. If there are more positive words then the text is positive. If there are more negative words then the text is negative. A threshold may be employed so that the difference must be above a certain level, providing the facility for a third "neutral" category.

The question arises as to how the word lexicons are compiled. Essentially this is a two stage process: First, candidate words should be harvested from a representative sample of the type of text the system is likely to encounter. Second these words should be annotated with word level sentiment values representing their contribution to the score.

There are typically two categories of method for ascribing word level sentiment score:

1. Human annotation: Each word is scored for how positive/negative it is considered to be by a human annotator. Several annotators may be deployed who each evaluate each word and the consensus of their scores taken. Obviously, this is a time consuming and costly process requiring skilled annotators. Care should be taken to ensure quality control between annotators. The accuracy of this method can be considered high.
2. Automatic annotation: Methods for automatically annotating lexicons have been developed. They usually involve measuring each word's proximity to a small set of hand picked seed words within a large corpus or machine thesaurus. These methods have the advantage of speed and can be repeated with different seed words to generate lexicons for different types of sentiment measurements. They have the disadvantage of low levels of word accuracy, with percentages in the low sixties in comparison with human word annotation.

These lexical methods of sentiment analysis are simple to construct and have good performance in terms of processing speed. Their accuracies tend to suffer because of the following reasons:

1. Lack of Word Sense Disambiguation (WSD). Many words have different meanings, or senses, depending on their context, and different word senses may have different sentiment scores. For example, in the phrase "the company was given a large fine" the word “fine” is synonymous with “penalty” and would generally be regarded as negative, however in the phrase "ACME is a fine company", "fine" would mean "good" and generally be considered positive. The act of deciding on which word sense is being used in a particular context is termed *word sense disambiguation*.

   Most lexicon based approaches do not have the means of addressing word sense disambiguation.
2. Linguistic Structure: Lexical approaches tend to ignore linguistic structure within text. Structure present in text means the sentiment expressed by the text can be very different from the sentiment values of individual words. A word's effect within a sentence is altered by those words around it. The simplest of these is negation, such as "this is a good idea" in comparison with "this is not a good idea". Both these sentences contain a single positive atomic feature "good" but their meaning is very different.

3.2 GRAMMATICAL PARSING

A heavy-weight approach to sentiment analysis uses traditional grammatical parse techniques to traverse a parse tree representation of the text applying sentiment rules to nodes. Typically, a parser operates on text one sentence at a time, applying (e.g. English) grammar rules to build a parse tree that represents the text as relations between words and phrases manifested as nodes in the tree. The leaf nodes represent words, and they can be ascribed sentiment scores in a similar way to lexical based approaches. The root of the tree is typically a node representing the sentence and so text is represented as a set of sentence trees.

A set of sentiment rules is used to traverse the tree by matching rules at each node, and bubbling up the sentiment scores for the branches to the root node. Thus complex linguistic constructs can be expressed as sentiment rules, modelling higher

sentiment bearing structure within text. For example, sentiment values might be ascribed to the subject and object of verbs differently: In the sentence "ACME disappoints the market", "ACME" is identified as the subject of the verb "disappoint" and "the market" would be the object. Thus ACME can be scored negatively, whilst "the market" is neutral.

English parsers have been developed over many years and are reasonably accurate. The trade-off however is with speed. Parsers tend to operate in the second timeframe, rather than the millisecond and so reduce their usefulness in time critical applications. Parsers for languages other than English are less common.

A further disadvantage of purely grammatical systems is that they tend to be brittle. That is, they can be sensitive to errors in parsing and rule-application: Small errors can have large effects further up the tree.

3.3 MACHINE LEARNING

A common machine learning approach to sentiment analysis is to treat it as a classification problem in the manner of standard topic classification. A typical approach is the so called bag-of-words method which uses a supervised learning algorithm to learn the mapping between the words that are present in a set of documents and a set of human annotated scores for those documents. Typically the process proceeds in the following manner: A large training set of example documents are classified by hand by humans; a feature vector for each training example is produced where each locus indicates the presence or absence of a particular word in the document; a suitable supervised machine learning algorithm is used to learn the mapping from feature vector to human sentiment classifications. An unseen document can then be classified by generating its feature vector and passing it through the learnt classifier.

The machine learning approach has the advantage that a lexicon of sentiment words does not need to be created: It is expected that the training algorithm will learn the predictive power of relevant sentiment words. Also classification should be relatively quick. However this method has a number of significant drawbacks outlined below.

3.3.1 Over-fitting

Over-fitting is a common pitfall of machine learning techniques. Over-fitting occurs when the training algorithm learns specific properties of the training set that are not present in the problem space in general and so the classifier poorly classifies unseen examples outside the training set. It can occur when the training set is too small, or the training set is not a good representation of the problem space. Specifically, in sentiment analysis, there are two aspects where machine learning techniques are prone to over-fitting:

1. *Large feature vector*: Typically the feature vector has a dimensionality in the 1000s or 10000s because of the number of words that are likely to be encountered in a reasonable sample of the target language. This means that the training set required to provide meaningful coverage of such a space must also be large. Manually annotating sufficient documents is a time-consuming and costly process and typically the space is under represented.

2. *Hitch-hiking*: The nature of language means there are more topic words than there are sentiment bearing words: Word occurrence histograms typically have very long tails, sothere are many uncommon words in any particular document and any sentiment signal is noisy. Without an explicit lexicon, the opportunistic nature of machine learning means it is likely to fixate on topic clues as predictors of sentiment rather than sentiment bearing words. This means it is easy for a training system to pick up on an inappropriate topic word as a good indicator of sentiment. For example, if a corpus contains a story about a "horrific bomb in a mosque" the word "mosque" is as likely to be incorrectly associated with negative sentiment as the word "horrific". We call this *hitch-hiking* and it is very difficult to design a corpus to avoid because sufficient positive and negative examples containing every word in the feature vector is required, leading to extremely large training sets.

The large feature vector and the issue of hitchhiking together with the dynamic nature of vocabulary, means that such machine learning-based sentiment systems require frequent retraining to maintain their effectiveness as vocabulary evolves. This is particularly significant in systems that process dynamic data sources such as news feeds.

3.3.2 Linguistic Structure

As with lexicon-based approaches to sentiment analysis, machine learning techniques that employ methods similar to bag-of-words suffer from the lack analysis of higher linguistic structure within the text. For example "it is not good, it's bad" will have the same feature vector as "it is not bad, it's good".

CHAPTER 4 THE THOMSON REUTERS APPROACH

Thomson Reuters's approach to sentiment analysis stems from the desire to build a system that is usable in the marketplace. A pragmatic approach has been taken to balance generalization, speed, and accuracy.

4.1 ARCHITECTURE

Thomson Reuters has recognised the strengths and weaknesses of other approaches and built a hybrid system to exploit the strengths and avoid the weaknesses. As such the sentiment engine has elements of lexicon matching, grammatical analysis and machine learning in it.

An architecture diagram is shown in Figure 1.

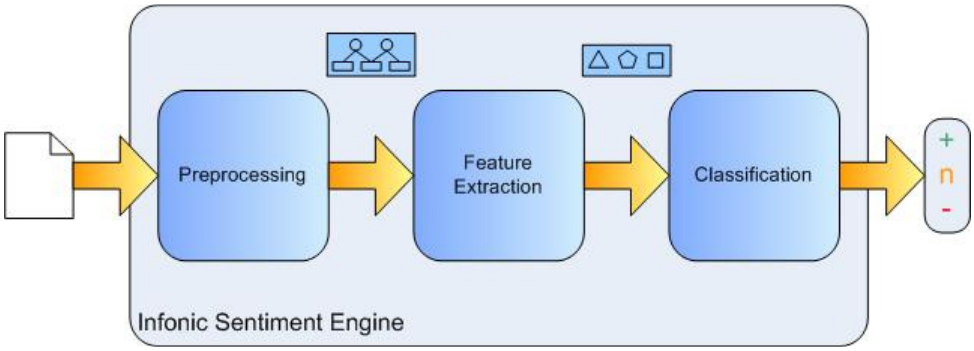

Figure 1: Sentiment Engine Architecture.

There are essentially three stages, *Linguistic Pre-processing*, *Feature Extraction* and *Classification*. These are described below.

4.2 LINGUISTIC PRE-PROCESSING

The purpose of the pre-processing stage is to transform the raw text into an internal representation that can then be mined for sentiment features by the feature extraction phase. The pre-processing uses components that would be familiar to anyone who has a background in computational linguistics. Essentially this stage performs the following processing pipeline:

1. *Sentence Splitting:* Divide the textual string into substrings representing sentences.
2. *Tokenisation:* Split each sentence into meaningful lexical tokens, which essentially equate to words and punctuation marks.
3. *Part of speech tagger:* Analyse each word within its context and decide its word-category: noun, verb, adjective, etc. For example, "bear" can be an adjective, noun or verb and so context clues are needed.
4. *Morphologically Stem:* Nouns and verbs are tagged with their base-form. That is, nouns are given their singular form and verbs their infinitive. E.g. "gone", "went" and "goes" are all recognised as being forms of the verb "go".
5. *Shallow Parse:* The sequence of tokens is split up into meaningful sub-sequences, or phrases, that take syntactic roles within each sentence. For example, in the sentences "the black cat slept" and "an unsuspecting electorate slept", the phrases "the black cat" and "an unsuspecting electorate" can both be considered to take the syntactic role of Noun Phrase and participate as the subject of the verb "sleep".

The pre-processing pipeline in stage 1 collectively performs a shallow parse of the raw text. In comparison with the traditional grammatical parse techniques described in Section 3.2, which can generate a multi-level tree-based analysis, the shallow parse produces a tree of only 2 levels. Its advantage as a technique is that it is very fast. The shallow parse forms the basis for the feature extraction phase.

4.3 FEATURE EXTRACTION

The feature extraction stage contains much of the sentiment processing and Thomson Reuters's related Intellectual Property. The goal of the feature extraction stage is to assemble sentiment features for each entity, which are then fed into the classification stage.

The feature extraction stage has two steps:

4.3.1 Lexical Analysis

In the lexical analysis step, the words (tokens) are assigned a set of atomic features. The features are obtained from a collection of hand annotated lexicons. There are roughly 16,000 words and 2500 phrases which have been triple annotated by human annotators. That is, each word has been looked at by three annotators and the consensus value taken. The lexicons contain the following word types:

Adjectives

Adverbs

Intensifiers

Nouns

Verbs

An example of a phrase in the lexicon would be "better than expected".

The exact nature of the annotation strategy is propriety, but each word is marked up with a number of features. To give a flavour they include: sentiment polarity, a means of dealing with word sense disambiguation, whether a word has a negating or intensifying effect on those around it and the effect a verb has on both subject and object.

The important thing to note is that the actual words are abstracted at this stage into a set of atomic sentiment features. For example, words such as "magnificent" and "fantastic" would have a similar representation, and for subsequent processing the actual word is not needed. The vocabulary of the system can be extended simply by adding words to the lexicons. The subsequent processing is therefore working on sentiment relevant abstractions of the words, the rules of grammar and Thomson Reuters sentiment patterns.

4.3.2 Sentiment Processing

After the words have been assigned atomic features, the second step in the Feature Extraction Phase consists of a traversal of the shallow parse tree to look for sentiment relevant structure within the tokens and the tree. This done using Thomson Reuters's sentiment patterns which analyse the tree structure to create a set of sentiment features for the text being processed.

Examples of the sort of processing that can be achieved are:

4.3.2.1 Negation

The most obvious sentiment relevant pattern within text is negation. The presence of words such as: “not”, “never”, “neither”, “against” with sentiment bearing words are recognised as negation, e.g. "not good". However, unlike simple rule-based systems, Thomson Reuters can handle cases that go beyond simple co-occurrence: "Things did not go very well for the company" will be recognised as negative.

4.3.2.2 Intensification

Some words take on different sentiment roles depending on their context, for instance intensifiers. Consider the role taken in by the word "terribly" in the following 2 sentences:

“The company is performing terribly.”

“The company is performing terribly well.”

In the first, "terribly" is acting as an adverb, whereas in the second it is acting as an intensifier of the word "well". The sentiment engine will score the first sentence negatively, and the second positively.

4.3.2.3 Verb Resolution

When scoring entities, the effect of a verb may vary depending on whether the entity of interest is the subject of the verb, or the object. For example:

> "ACME disappointed the market."

This sentence will be scored negatively for ACME and neutral for "the market". This is an example of the active verb use, similarly in the passive phrase:

> "The market was disappointed by ACME."

ACME will again be scored negatively and "the market" scored neutrally.

This gives a flavour of the sorts of processing that is occurring in the feature extraction module. Ultimately, the lexical analysis and sentiment patterns produce a feature vector representing the sentiment features for the entity being scored. The feature vector is used as input to the third stage in the engine, the classifier.

4.4 CLASSIFICATION

The job of the classification stage is to produce the final sentiment score for the text from the features. Recall in Section 3.1 where we discussed the threshold method in the lexicon approach for picking a classification based on the number of positive and negative features. We are presented with a similar problem: that of deciding the final classification, given the analysed features. However, rather than using a somewhat arbitrary threshold, we use machine learning techniques to learn the mapping from features to the classification of real data.

The classifier is a simple three layer back-propagation neural network with weight relaxation. Because we are using the feature representation of the text rather than the words, the dimensionality is low. It is trained using 5000 triple annotated news articles spanning the 14 months from Dec 2004 to Jan 2006.

4.4.1 Output

The classifier produces three real-valued outputs between 0.0 and 1.0: They are the probabilities that the current input is classified as positive, neutral or negative. Since they are probabilities, the outputs sum to 1.0. They may be used as confidence measures and thresholded to provide precision/recall tradeoffs.

4.5 CLASSIFICATION ACCURACY

The sentiment engine has been tested on thousands of news articles covering a variety of topics, from small newswire alerts to longer editorial pieces. On average, the system achieves around 75% accuracy against the average assessment of human analysts. This is impressive when one considers that human beings only agree with each other on any given article around 82% of the time. And, it is important to note that the system processes around 10 articles a second whereas a human processes around 6 articles an hour.

A graph of precision verses recall is shown in Figure 2.

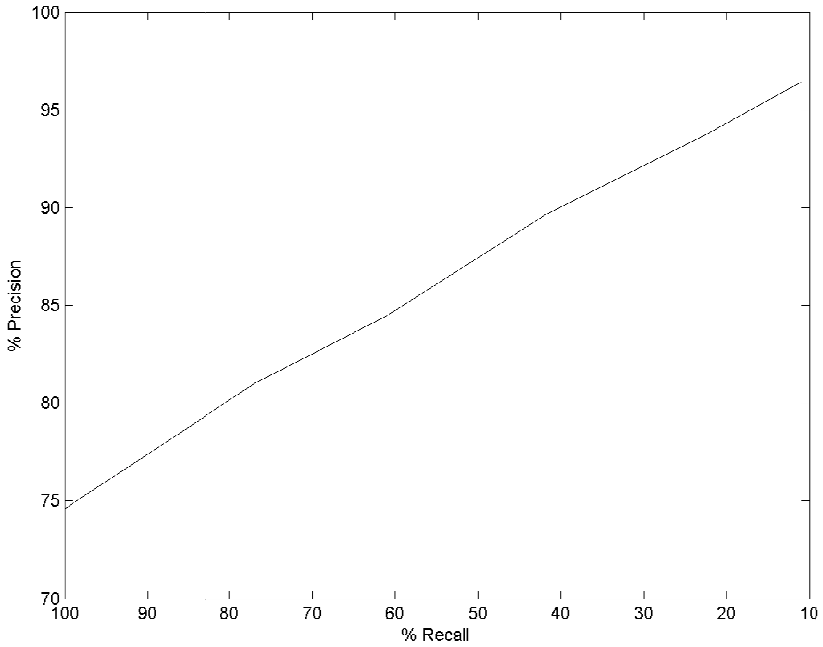

Figure 2: Precision versus recall.

In the Figure we can see what happens as we vary the threshold level at which the system makes a prediction, as described in the previous section. The X axis displays the percentage of articles that the system makes a prediction on, and the Y axis display the accuracy of these predictions. So for example:

- When the system is forced to make a prediction on 100% of articles it gets around 75% correct.
- If the confidence level is set so that the system only makes predictions for 50% of articles then around 87% of these predictions will be correct.
- If the confidence level is set so that the system only makes predictions for 20% of articles then around 95% of these predictions will be correct.

CHAPTER 5 DISCUSSION OF DESIGN ISSUES

This section describes the various design challenges that have been addressed by the TRNA system.

5.1 LEXICAL GENERATION

The lexicons used in the News Analytics engine are machine harvested and human annotated. That is, a large corpus of sample documents from the same source as those to be classified is collected and the words from each lexical word type are extracted, i.e. adjectives, adverbs, intensifiers, nouns, verbs. A lexicon is selected from the most common words in each category, and those are passed for human annotation. Each word is annotated by three humans, who are working from a carefully prepared script. The words are presented to each annotator in a random order and the consensus taken as the final word score.

For News Analytics, the corpus of sample documents contained 200,000 Reuters News articles, randomly selected from 2004-2006. We have performed lexical coverage tests where we extracted the lexical words from two distinct years: 2003 and 2005 and measured the overlap to be 90%.

The method of hand annotating the Lexicon is a fairly time consuming and costly process, but it is felt that this method ensures a high quality resource and produces high accuracy classification. The Lexicons currently contain 16,000 words which it is stressed are sentiment bearing words: Words that contain no sentiment have been pruned e.g. "table", "turquoise". It has been shown this is sufficient to provide good coverage of unseen data with the lexical coverage tests.

5.2 WORD SENSE DISAMBIGUATION

As discussed above, the issue of word sense disambiguation is a big problem in sentiment analysis. Words can have a number of senses, which can take different sentiment values. It is tackled though the News Analytics lexicon annotation scheme and sentiment patterns. As an example of the types of distinction the News Analytics engine can handle, consider the different uses of the following words in the contexts shown:

- "terribly"
  “ACME is performing terribly.”
  “ACME is performing terribly well.”

- "fine"
  “ACME received a large fine"
  “ACME sells fine sand and aggregates”
  “ACME is a fine company”

The News Analytics engine's ability to handle cases like these distinguishes it from lexical and machine learning techniques.

5.3 SENTIMENT SIGNIFICANT LINGUISTIC STRUCTURE

The Thomson Reuters hybrid approach enables the capture of sentiment relevant linguistic structure within text, in contrast to pure lexicon and machine learning approaches. This is implemented through the use of our shallow parse and sentiment pattern recognition algorithms, as described above. These techniques perform a similar task to the full grammatical parse/rule based techniques described in Section 3.2 and are sophisticated enough to allow significant utilization of sentiment relevant structure. They also have the advantages of efficiency, in that they very fast and scale well with document length. Likewise, they avoid the brittle nature of a full parse, giving robust results that are resilient to the error levels that are seen in other sub-modules in the pipeline.

5.4 FINAL CLASSIFICATION

When the threshold method is used to determine the final classification, as in the simple lexicon method, there is an assumption that the output classification has a simple relationship to the number of each type of feature. The machine learning technique deployed by Thomson Reuters ensures that the mapping from sentiment features to the assigned score is modelled on real data. This has the advantage that the mapping is the best one available, given the available information. Despite the relative simplicity of the machine learning architecture used, this method has given significant improvement over threshold methods.

5.5 OVER-FITTING

A big worry with any type of classification is that of over-fitting. This is where the system in question performs well on data within the annotated training data (*in-sample data)*, but gives disappointing results on *out-of-sample* data. In particular, pure bag-of-words machine learning methods used in Computational Linguistics are especially prone to this, given the high dimensionality of their feature vectors, the relatively low number of training points and the noisy nature of linguistic data. The Thomson Reuters approach has been developed with a particular care to avoid over-fitting. The measures we have in place include:

1. The lexicons are hand annotated rather than learnt, reducing noise.
2. The lexicons are harvested from 200,000 financial news articles spanning 2004-2006, i.e., hugely more than the classifier's training set.
3. The words themselves are abstracted using our annotation scheme, vastly reducing the dimensionality and data spikiness.
4. The linguistic rules are rules of the language, i.e. English, which change infrequently.
5. The feature vector is relatively short, allowing a realistic training set size.
6. The training algorithm uses a weight relaxation factor to avoid over-fitting.

5.5.1 Linguistic Sub-Modules

It should be noted that the complex nature of language means that no computational linguistics technology is 100% accurate. Further, advanced systems are created by pipelining functional sub-modules, and there is a danger that errors will multiply to swamp any useful analysis signal. It becomes a significant effort to manage the inaccuracies of even state-of-the-art sub-modules within a system and Thomson Reuters has been very mindful of this during the design of our sentiment system.

5.5.2 Validation Tests

We have performed 10-fold cross-validation tests to measure out-of-sample performance. That is, the annotated set of 5000 articles, spanning 14 months, is randomly divided into 10 equally sized chunks, or *folds*. Ten train/test cycles are performed where each fold is removed from the training set and used as a test set. That is, the system is trained on 9/10th and tested on 1/10 ten times so that every article gets to be a member of both training and test sets. The performance on the 10 training sets and 10 test sets is measured and averaged and we can report a <2% drop in out-of-sample accuracy when compared with in-sample accuracy, on average. This demonstrates satisfactory generalization ability.

5.6 SPEED

A major consideration during the design of the News Analytics engine was to produce a system that was usable in the marketplace. Processing speed is an important consideration. The pure Lexicon and Machine Leaning techniques have the advantages that they are quick and scale well, but traditional grammatical parse techniques are generally slow and have poor worst-case performance.

The Thomson Reuters solution is able to provide accuracy with good performance. It scales linearly with article size and has good worst-case performance. This is largely down to the use of a shallow parse to perform the linguistic processing rather than a full grammatical parse of English. The shallow parse is a single pass algorithm that is well behaved.

CHAPTER 6 RELEVANCE

The scores produced for a company by the sentiment engine are probabilities that the analysed article is positive, negative or neutral about that company. The engine can score articles of all sizes, from single line news feed alerts to large opinion pieces, and scores are produced for each company that is mentioned in the article. In TRNA, a list of companies of interest is supplied in the metadata that accompanies the article and this is used to determine which companies the article is scored for.

It is inevitable in news articles that some companies are talked about more than others. A news report may be about a particular event that involves a single company, such as press release, earnings announcement or threat of redundancies. There may be two or more predominant companies, such as in a merger or acquisition announcement or a market piece about two rivals. Alternatively an article may have substantive text about a number of companies, such as a news roundup or market analysis piece. Conversely, a company may only be mentioned in passing in a piece largely about other companies.

It may be useful to distinguish between these cases when consuming the sentiment scores. It may be desirable to take particular notice of scores that are generated for a company from an article that is largely about that company. On the other hand, it may be desirable to give less weight to companies that are passing mentions. The sentiment scores in themselves don't convey these distinctions, so the News Analytics engine has a relevance factor is published for each company alongside the sentiment scores.

The aim of the relevance factor is to distinguish between these cases:

1. The article is predominantly about the company.
2. The company is one of many mentioned in the article.
3. The company is a passing mention in the article.

The relevance is an occurrence-based measure, calculated from the number of mentions a company has in comparison to other companies. That is, it does not make judgements on the content or topics mentioned within the article.

6.1 RELEVANCE CALCULATION

The relevance score for a company is calculated from the number of times it is mentioned in the text compared to other companies and organizations.

Within the relevance module, there are 2 sub-modules whose job it is to find the occurrences of companies and organizations:

1. The *synonym matcher* finds the occurrences of the company that is currently being scored. Because these are known companies, the system has metadata about them including synonym data about how they may appear in text. For example, for the company IBM, there may be the synonyms "IBM", "I.B.M.", "International Business Machines" and "Big Blue". The synonym matcher is used to find all occurrences of each scored company both the article body and the headline of the article.
2. The *named entity detector* finds the mentions of all the other companies and organizations that are discussed in the text. These can be considered unknown companies and the detector uses word morphology and contextual clues to recognise company mentions, which are co-referenced to collect together all the occurrences of each company. For example, it might find the strings: "Microsoft Corp", "IBM", "International Business Machines" and "Microsoft" and infer that "IBM" and "International Business Machines" refer to the same company, as do "Microsoft Corp" and "Microsoft". The number of mentions of each company is counted: In this case there are 2 companies with 2 mentions each.

The relevance is then calculated using the number of mentions of the scored company found with the synonym matcher compared to the number of mentions of the most common named entity company and the total number of company mentions

found using the named entity detector.

This calculation gives the relevance a value between 0.0 and 1.0 for each scored company. In addition, if the company of interest is mentioned in the article headline, then it receives a relevance of exactly 1.0.
This measure successfully allows distinction between the three cases:

1. When the relevance equals, or is close to 1.0, the company is one of the predominant players in the article.
2. When the relevance is between about 0.8 and 0.2, the company is one of several mentioned substantively in the article.
3. When the relevance is less than about 0.2, the company is a minor player or a passing mention.

It should be noted that the mentioned companies' relevance scores are not normalized. They do not sum to 1.0 across companies. There may be more than one company for which a particular story is relevant: For example, if 2 companies are mentioned in the title of an article discussing their merger, both companies will receive a relevance of 1.0.

The relevance calculation provides a means of filtering the sentiment scores to give weight to predominant articles and reduce the noise created by passing mentions. It is also a tool in the armoury for estimating the volatility of signal for a particular company: A period where there are many highly relevant stories about a company indicates that an important event may be occurring with resultant effects on the volatility.

CHAPTER 7 NOVELTY

The sentiment and relevance scores produced by the News Analytics engine deal with the content of a news story. Together they convert the news into numeric polarity and magnitude measurements for the use as indicators. Since every relevant news item generates a set of scores on a news tick basis, news volume information is available from the frequency of ticks for each company.

It is the nature of news stories that they have a lifetime: They occur at a moment in time and evolve over time. Each story may be reported as a series of news items or articles. There tends to be an immediate response to a newsworthy event resulting in a news alert, or breaking-news item followed by a first take, then subsequent updates to the story. In addition there are reaction pieces, analysis articles and roundup articles that expand on the story as time progresses.
It is obvious that new news is important for market insight. *Breaking news* implies an underlying event has occurred in the world, which may move or report a move in share price. So this further implies that breaking news is an indicator of volatility. In a given news feed, there are likely to be multiple stories running simultaneously. Determining which articles represent stories that are breaking and which are parts of stories that have been rumbling along for a while would give a handle on the volatility.

In News Analytics, Thomson Reuters has used a news content clustering technology to provide such a "breaking news detector" in the form of the *novelty score*.

7.1 NOVELTY CALCULATION

The novelty score determines the *uniqueness* of a news article in comparison to those that have been seen before, within a set of configurable *history periods*. The News Analytics engine keeps a cache of metadata on all processed articles for the duration of the longest history period in a database. The metadata is a vector representation of the article content, like a linguistic fingerprint, that represents the whole article. When processing an article for a company, the News Analytics engine retrieves the metadata of previous articles that mention the company and compares the current article with all the retrieved articles to get a *distance* measure between them. The distance measure is used in a *similarity function* to determine which of the retrieved articles the current article is similar enough to, to declare that they are *linked articles*. That is, two articles are linked if they are similar, and not linked if they are not similar. The similarity function is more sophisticated than a simple threshold because it has to determine whether two articles of differing length are linked or not — a common issue in similarity measures of textual data.

Once the set of linked articles has been calculated, they are sorted in to five buckets according to age, each bucket corresponding to a history period. The default history periods are 12 hours, 24 hours, 3 days 5, days and 7 days, highlighting different timescales. News Analytics publishes the number of linked articles in each of the five history periods, called the *linked counts*, and a list of up to ten item ids, called the *linked ids*. If there are more than ten linked items, the linked ids of the first five and the last five by timestamp are published.

The novelty scores, comprising the linked counts and the linked ids, give a measure of whether an article is breaking news or not. If the linked counts in the history periods are zero, the article is breaking news: There has not been a similar article for the duration of the history period. Likewise if the linked counts have a high value, then there have been lots of similar articles during the history period, and the news can be considered known to the readership.

The use of this linguistic measure of article similarity, allows News Analytics to determine if two items are talking about the same real world events outside the usual Reuters story structure conveyed by the Reuters metadata that accompanies the scores.

CHAPTER 8 SUMMARY

The Thomson Reuters technology within TRNA has produced a state-of-the-art text analysis tool comprising the sentiment, relevance and novelty scores. These tools allow the algorithmic trader to get a handle on the polarity, magnitude, volume and volatility of the market–moving news signal. They perform analysis at the company level, providing per–equity indicators.

The sentiment engine is a hybrid system that uses elements of lexical analysis, linguistic parsing and machine learning techniques, utilizing the advantages and avoiding the pitfalls of each. It has been carefully designed to produce accurate scoring with good performance and generalization capability. Particular attention has been given to avoid over-fitting and the system is accurate on unseen articles.

In TRNA, Thomson Reuters technology is being utilized in a ground–breaking application to turn unstructured news data into structured indicators for algorithmic trading.

For more information

Send us a sales enquiry at
forms.thomsonreuters.com/qed/

Read more about our products at
thomsonreuters.com/products_services

Find out how to contact your local office

thomsonreuters.com/about/locations

Document Version 1.1
Date of issue: February 2013

THOMSON REUTERS™