Sentiments on a Grid: Analysis of Streaming News and Views

Khurshid Ahmad[1], Lee Gillam[2], David Cheng[2]

[1] Department of Computer Science
Trinity College, Dublin
khurshid.ahmad@tcd.ie

[2]Department of Computing
University of Surrey
{l.gillam, d.cheng}@surrey.ac.uk

Abstract

In this paper we report on constructing a finite state automaton comprising automatically extracted terminology and significant collocation patterns from a training corpus of specialist news (Reuters Financial News). The automaton can be used to unambiguously identify sentiment-bearing words that might be able to make or break people, companies, perhaps even governments. The paper presents the emerging face of corpus linguistics where a corpus is used to bootstrap both the terminology and the significant meaning bearing patterns from the corpus. Much of the current content analysis software systems require a human coder to eyeball terms and sentiment words. Such an approach might yield very good quality results on small text collections but when confronted with a 40-50 million word corpus such an approach does not scale, and a large-scale computer-based approach is required. We report on the use of Grid computing technologies and techniques to cope with this analysis.

1. Introduction

Corpus builders are creating giga-word corpora, some using the web as a resource and others using the typical fodder of many corpora – the news wire. Information extraction folk are also looking at collections of texts, some in the form of a systematically organised corpus, for identifying seminal events, important persons and so on. The diverse mixture of texts - news wire corpora contain political, financial, sports and other news - should make the extraction of meaningful information a difficult task.

Analysis of specialist text corpora, of financial news or political news or sports news, invariably starts with a pre-determined vocabulary, and is often accompanied with a gazeteer of place names and a directory of personal names. Such an approach may suit synchronic analysis of texts or that of a text archive. However, in a real-world specialist news stream there is coverage of new devices, goods, and services, on the one hand, and the names of new people and of new companies that may or may not be in the gazeteers or the directories. The analysis of streaming news texts has been bestowed a strategic importance due to economic globalisation on the one hand and the global nature of terrorist threats on the other. News and views, it appears, can move the world: it can make or break people, companies, perhaps even governments. At time of writing, the actions of a Danish newspaper and follow-up stories by other international news agencies have led to international civil unrest; news and views about Brazilian government bonds expressed by financial traders in London, through news wire and inter/intra- office emails has significant impact on London trading and on the Brazilian economy (Hardie and Mackenzie 2005); a $300 million New York hedge-fund firm collapsed following the circulation of an email that expressed doubts about the validity of the operation in the Russian Republic (Mackenzie 2000).

There are a number of new disciplines that incorporate the effect of news and views expressed by, and sometimes not expressed or deliberately concealed by, investors and traders into the economic and financial analysis of financial instruments (currencies, shares, derivatives: Engle and Ng 1993) an that of firms (Baker and Wurgler 2004). The new disciplines are referred to as investor psychology and behavioural economics and are known more broadly as economic sociology. A principal method in these disciplines is called news impact analysis: pioneered by Engle in 1977 for studying the correlation of negative/positive news announcement on British inflation, this method is now used extensively in econometrics. The term *news impact* is used variously in the investor psychology / behavioural economics literature: for Engle and Ng it was the collocation of significant changes in the value of a financial instrument – the so-called volatility – expressed as a change in the structure of a financial time series – with the timing of the announcement from the various financial and regulatory agencies that could be used to infer whether the sentiment was negative or positive. The negative news, according to Engle's Nobel Lecture in 2003, has a more persistent effect than that of positive news. Other authors have extended Engle's news impact analysis by doing a limited content analysis on pre-selected keywords and sentiment words (Andersen et al. 2002, DeGennaro and Shrieves 1997). More recently, neural networks have been trained to associate news stories with market movements with the intention of predicting market movement based on the contents of news items (Koppel and Schtrimberg 2004)

In this paper we report on constructing a finite state automaton comprising automatically extracted terminology and significant collocation patterns from a training corpus of specialist news (Reuters Financial News) that in turn can be used to unambiguously identify sentiment-bearing words that might be able to make or break people, companies, perhaps even governments. The volume of news arriving (c. 1.5 million words per weekday and 0.5 million at weekends) that is to be analysed, in addition to the analysis needed for constructing the automaton based on archives containing last week's, last month's or last year's news, depending on the accuracy desired, creates a significant computational resource requirement. The work was partially carried out in a UK e-Science project sponsored by the Economic and Social Sciences Research Council (ESRC). We created a Grid of 24 machines (c. 80 processors) for parallel news analysis. Standard Grid middleware was used, including Globus Toolkit and Condor, and we have shown how the throughput increase decays with increasing computational power. The parallel processing of texts is essential due to the quantities of data involved. From as much as 0.5 million tokens arriving per hour at peak times, on average we extract just under 5000 potentially sentiment bearing sentences. These sentences are processed using the finite state automaton for disambiguation, with an approximate yield of around 50-100 sentences that are truly sentiment bearing.

The paper presents the emerging face of corpus linguistics where a corpus is used to bootstrap both the terminology and the significant meaning bearing patterns from the corpus. Our research shows that much of the current content analysis software systems require a human coder to eyeball terms and sentiment words. Such an approach might yield very good quality results on small text collections but when confronted with a 40-50 million word corpus a large-scale computer-based approach is required. A key goal is to avoid the production of false positives while processing these kinds of text volumes in as short a time as possible.

2. Background

Assessing attitudes to a range of artifacts, including films and cars, banking institutions, and holiday destinations (Turney 2002) has been referred to as sentiment analysis or opinion analysis, or affect analysis or opinion mining (Grefenstette et al., 2004). Researchers have trained classifiers on corpora of movie reviews (Pang et al., 2004, Pang et al., 2002, Bai et al., 2005), and used sentiment extraction for reviews of music and digital cameras (Yi et al., 2003). Sentiment analysis techniques may be useful for Customer Relationship Management (Roussinov et al., 2003) and identifying abusive postings (or flames) in Internet newsgroups (Spertus 1997). More socially beneficial applications include measuring the “reassurance gap”, the difference between crime rates and the public perception of crime (Fielding 1995, Fielding, Innes and Fielding, 2002), and discovering “internal war” (Kaldor 1999), where conflicts may have a historical basis and rely on collective memory or be based on a reinvention of identities.

Product sentiment analysis tends to be quite focussed. The domain fixed, but in large part the item about which the sentiment is being expressed has also been fixed – it is the movie, or the product, or the company. There may be some deviations from the item – comparisons and so forth. Durbin et al., (2003) differentiate between “analytic methods (e.g. named entity extraction) that provide specific items of information, and synthetic methods (e.g. topic identification) that provide a global characterization”. According to these authors, much of the current work on sentiment analysis, opinion analysis and affective rating fits into this second category. Financial sentiment, however, does not seem to fit such a simple distinction. Financial news may be about a single company, but may also be identifying sentiments about other companies in the same sector. It may contain sentiment about an item such as oil prices rising that have a negative impact on oil-related industries such as the aviation industry but perhaps have a positive impact on the profits of petroleum companies: oil prices rise, Shell profits rise, BA profits fall. It may contain positives but with a negative outlook. We need to be able to accurately identify the *thing* about which we are extracting the sentiment, and at the same time, at a more broad level, we desire the global characterization for, for example, a company, an industry sector, a price index or a currency, or any number of these. Extracting financial sentiment, then, appears to demand a hybrid approach.

Turney’s approach to sentiment analysis seeks the “Semantic Orientation”: all phrases matching manually selected part-of-speech patterns are scored according to their proximity to either of the words “excellent” or “poor” in a proxy for general language – the AltaVista search engine. Turney claims classification accuracy of bank and car reviews of 80-84% based on 410 articles, but has difficulty classifying movie reviews: his approach classifies the full text as positive or negative, and for movie reviews “the whole is not necessarily the sum of the parts”. This is a useful benchmark experiment, however the scientific repeatability is problematic: firstly, the additional indexing of information by a search engine is likely to change the scoring over time; secondly, this particular experiment cannot now be repeated due to changes to the function of the AltaVista Engine.

3. Method

Financial investors, it appears, combine information from a variety of sources and may pay more or less heed to certain sources at certain times and ignore others. Calendar events such as announcements of economic indicators can be of greater importance than the financial time series. The motivation for our work is to discover when and how news texts impact on financial decisions: quantifying their effects in the so-called Efficient Market Hypothesis. The result of the impact should be measurable in unexpected or unpredictable movements of financial instruments.

In the methods of sentiment analysis discussed, either the sentiment ‘variables’ and metrices use information proxies, or they rely on pre-selected keywords and phrases – the best guesses or intuitions of the researchers. These methods are designed to avoid, and perhaps ignore, ambiguity that is inherent in natural language based communication. However, it is important to explore whether or not sentiment-bearing phrases can be extracted with a minimum of ambiguity, where the premium is on avoiding false positives, without relying on prior knowledge.

The method uses relatively large collections of texts: a training collection in the specialist domain and a representative general language (reference) corpus. We automatically discover domain-specific keywords and build statistically relevant collocation patterns using these keywords that may contain ontological statements or sentiments. Once this set of patterns has been manually validated, it can be used as the basis for sentiment extraction. Over time, the collection of patterns may evolve by discovering new patterns from the incoming texts. The method does not rely on any overt access to an external knowledge base.

We start with a contrastive analysis of a training collection in the specialist domain and the representative general language (reference) corpus. Reuters Corpus Volume 1 (RCV1) and the British National Corpus (Aston and Burnard, 1998) are used, and we extract key words automatically using a "smoothed" weirdness calculation (Gillam and Ahmad 2005 - MLDM). Grammatical words (the, a, an, and, but..), usually described as a stop list, have a very similar distribution (contrastive values close to 1), but subject specific words have a different distribution (see Table 1):

Word	Weirdness
percent	157.84
market	8.49
company	5.09
bank	10.99
shares	19.51

Table 1. Most "weird" words: "percent" occurs 157 times more frequently in RCV1 than in BNC

Next, we consider that collocation patterns, combinations of words that occur together frequently, are indicators of meaning and intent of the author, and these weird words will likely be used in the most meaning-bearing collocations (Gillam 2004). We use a stoplist and compute both contiguous and non-contiguous statistically significant collocates, 5 words either side of the weird words. Key collocates of *percent* are *up, rose, rise, down* and *fell.* This results in contiguous patterns *rose X percent, X percent rise*, and the discontiguous *up [by] X percent.* Our method then computes collocates of these collocates. The collocation patterns suggest that the metaphorical words, *rose, fell, up, down*, usually used to refer to movement of objects in physical space have been transferred (the origin of the word metaphor) over to the change in the value of the rather abstract financial instruments.

The finite state automaton (see Figure 1) is derived using the percentage of occurrences of dominant collocation patterns in the set of all collocations of the individual word. Validation of these patterns as ontological or sentiment-bearing enables their use against incoming texts to quantify, for our purposes, the amount of sentiment occurring at a given time such that it can be aligned with financial time series (Gillam and Ahmad 2006 - GiF). These frequent collocates have an unambiguous interpretation, and the avoidance of ambiguity is the cornerstone of modern information retrieval. The frequent collocates of collocates have still more unambiguous interpretation.

4. Real-time analysis and Evaluation

The processing of large volumes of texts, to derive the automaton, and later to process the volumes of incoming texts, is a computationally intensive task. We use a Grid of 24 machines, and our incoming news is from a financial datafeed provided by Reuters Financial Services (Gillam, Ahmad and Dear 2005) with a current peak flow of over 10,000 news items per day. To obtain real-time analysis, we have experimented in configurations of 1 to 64 CPUs – with a mere gain of 7% in moving from a 48-CPU grid to a 64-CPU grid configuration. With the forthcoming procurement of a further 100+ processors and terascale storage at Surrey, and access to 2000+ processors of the UK's e-Science National Grid Service[1], performance degradation and optimization will be future considerations.

Using this capability, we can identify between 1,000 and 10,000 sentences that may indicate sentiment based on the keyword alone, in a corpus of between 10,000 to 100,000 tokens arriving per hour. The disambiguating power of the local grammar patterns results in these 1,000 to 10,000 being filtered down to the 'true' sentiment bearing sentences - 10 to 100 (figures are rough approximations). 99%, then, of indicatively positive or negative sentiment words are not being used in the discovered positive or negative contexts (false positives, and, indeed, false negatives, in a very real sense), and are filtered out.

Focus on dominant collocations from 1996-7 may miss newer results due to, for example, changes in editorial policy. The degree of filtering achieved suggests that ambiguity may be a problem in financial texts, but further investigation would be needed to confirm this. Ambiguities in language typically occur because a pivotal verb (or noun) in a sentence can be replaced by other verbs (or nouns). The specialist nature of financial news suggests a restriction of the majority of such verbs (nouns) to a small subset of such words in the language and thereby minimizing ambiguity. Such values for filtering figures may be skewed by, for example, many instances of Michael *Rose* being removed – but recall use of the stoplist in the pattern construction. The approach presented is contrary to the current paradigm of natural language processing that is grounded in *universal grammar* – where many words can be used interchangeably. The approach used is called *local grammar.*

The capabilities of Grid technologies have helped us to write and test the programs that recognize the patterns automatically, and to be able to better verify and validate the results of the experiments, with repeatability of the experiment to the fore. This degree of throughput capability is important for dealing with the deluge of texts from one source alone; further work will expand efforts to the consideration of multiple competing and co-operating (syndicating) sources where there will be further challenges, especially for sentiment analysis: the ontology – *what there is* – may maintain some consistency, but the sentiment – *what people think about what there is* – necessitates efforts.

Acknowledgements

This work was supported in part by research projects sponsored by the EU (LIRICS: eContent-22236) and by the UK Economic and Social Research Council (ESRC) (FINGRID: RES-149-25-0028) under the e-Science programme.

[1] See: www.ngs.ac.uk for further details

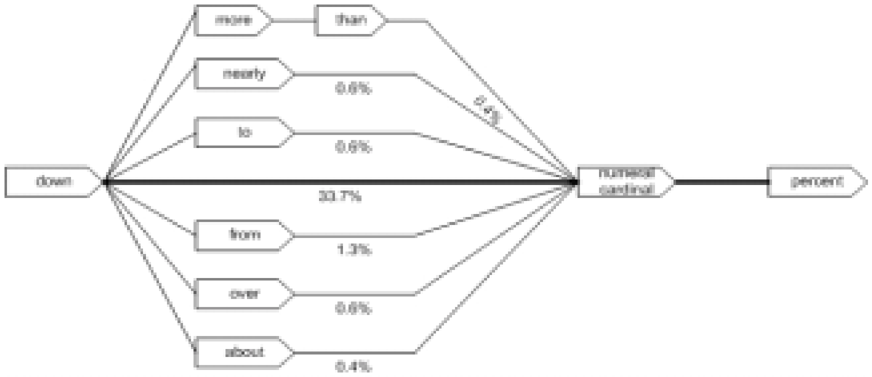

Figure 1: A finite state automaton for recognising negative sentiment sentences comprising ‘down’

References

Andersen, T. G., Bollerslev, T., Diebold, F X., Vega, C. (2000) “Micro effects of macro announcements: Real time price discovery in foreign exchange”. National Bureau of Economic Research Working Paper 8959, http://www.nber.org/papers/w895

Aston, G., Burnard, L. (1998) “The BNC Handbook”. Edinburgh: Edinburgh University Press

Bai; X., Padman, R., Airoldi, E. (2005) “On Learning Parsimonious Models for Extracting Consumer Opinions”. Proc. of HICSS '05. IEEE press.

Baker, M., and Wurgler, J. (2004) “Investor Sentiment and the Cross-Section of Stock Return”. NBER Working Papers 10449, Cambridge, Mass National Bureau of Economic Research, Inc.

DeGennaro, R., Shrieves, R. (1997) “Public information releases, private information arrival and volatility in the foreign exchange market”. Journal of Empirical Finance vol. 4, p 295–315.

Durbin, S.D., Richter, J. N. Warner, D. (2003) “A system for affective rating of texts”. In Proc. of 3rd Workshop on Operational Text Classification Systems (OTC-03), Washington, DC.

Engle, R. F., Ng, V. K. (1993) “Measuring and testing the impact of news on volatility”. Journal of Finance. Vol. 48, p. 1749-1777.

Fielding N. (1995). “Community Policing”. Oxford: Oxford University Press

Fielding, N., Innes, M., & Fielding, J. (2002). “Reassurance Policing and the Visual Environmental Crime Audit in Surrey Police: a Report”. Guildford: Univ. of Surrey Department of Sociology.

Gillam, L. and Ahmad, K. (2005). “Pattern mining across domain-specific text collections”. LNAI 3587, pp 570-579

Gillam, L. and Ahmad, K. (2006) “Financial data tombs and nurseries: A grid-based text and ontological analysis”. Proc. of 1st Intl. Workshop on Grid Technology for Financial Modeling and Simulation (Grid in Finance 2006).

Gillam, L., Ahmad, K. and Dear, G. (2005). “Grid-enabling Social Scientists: some infrastructure issues”. Proc. of 1st International e-Social Science Conference

Gillam, L. (2004) “Systems of concepts and their extraction from text”. Unpublished PhD thesis, University of Surrey.

Grefenstette, G., Qu, Y., Shanahan, J.G., Evans, D.A. (2004) “Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application”. Proc. of Recherche d'Information Assistée par Ordinateur (RIAO).

Hardie, I., and MacKenzie, D. (2005) “An Economy of Calculation: Agencement and Distributed Cognition in a Hedge Fund”. http://www.sps.ed.ac.uk/staff/An%20Economy%20of%20Calculation.pdf

Kaldor, M. (1999). New and Old Wars: Organised Violnece in a Global Era. Polity Press: Cambridge.

Koppel, M., Shtrimberg, I. (2004) “Good News or Bad News? Let the Market Decide”. AAAI Spring Symposium on Exploring Attitude and Affect in Text. Palo Alto, 2004, p. 86-88. : AAAI Press.

Mackenzie, D. (2000). “Fear in the Markets”. London Review of Books. Vol 22 (No. 8).

Pang, B., Lee, L. (2004) “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”. Proceedings of the ACL, p. 271-278

Pang, B., Lee, L. (2002) Vaithyanathan, S.. “Thumbs up? Sentiment Classification using Machine Learning Techniques”. Proc. of EMNLP, pp79-86.

Roussinov, D., Zhao, J.L. (2003) “Message Sense Maker: engineering a tool set for customer relationship management” Proc. of 36th Annual Hawaii International Conference on System Sciences (HICSS).

Spertus, E. (1997) “Smokey: Automatic recognition of hostile messages”. Proc. of Conference on Innovative Applications of Artificial Intelligence, Menlo Park, CA: AAAI Press. p. 1058-1065.

Turney, P.D. (2002) “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”. Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, p. 417-424.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003) “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques”. Third IEEE International Conference on Data Mining (ICDM), p. 427 – 434. IEEE press.