Textual and Quantitative Analysis: Towards a new, e-mediated Social Science

Khurshid Ahmad, Lee Gillam, and David Cheng

Centre for Knowledge Management, Department of Computing, University of Surrey,

Guildford, Surrey. UNITED KINGDOM

[k.ahmad, l.gillam, d.cheng}@surrey.ac.uk

Introduction

The analysis of large volumes of quantitative data in social sciences has a long tradition: mathematical modelling and statistical analysis can now be performed on large, distributed data sets and the results of such computations can be variously visualised. Qualitative research, and the concomitant data acquired, is no longer confined to 'issues of "subjective" meaning' but embraces exciting areas of 'language, representation and social organisation' (Silverman 2004:1). Research in qualitative analysis deals with patterns of linguistic discourse, with researchers scrutinizing linguistic output obtained through interviews, focus-group sessions, e-mail archives, and so on. The discourse patterns are often identified and 'coded' by hand for subsequent retrieval and analysis.

The conjunctive analysis of quantitative and qualitative data in social sciences is motivated by the availability of large volumes of such data and by the desire of the researchers to systematically include *lived experience* of individuals in their analysis. Consider three areas where such a conjunctive analysis may bear fruit: (i) investor/trader *credulity* in financial markets (Simon 1992, Kahneman 2002); (ii) the *reassurance gap* in policing (Fielding, Innes and Fielding, 2002); (iii) *totalising war discourse* that leads to ethnic/racial conflicts (Jackson 2004:22). In these three examples, results of the analysis of quantitative data may be at variance with the results of a qualitative analysis: (i') a company's fundamentals – its book value – can be at substantial variance with its share value due to *market sentiment* (Baker and Wurgler 2004); (ii') falling crime rates do not explain the increasing sense of insecurity amongst British citizens (Dalgleish and Myhill 2004); and, (iii') disinformation, against groups of people, leading to 'internal conflicts' (Kaldor 1999, Jackson 2004).

Much of the investor credulity, the reassurance gap, and the totalising war discourse is disseminated in written texts, including newspaper reports, web-sites, books and magazines, and these texts in some sense create a 'new social reality' (*cf*. Markham 2004). Content analysis, especially analysis of written texts (Mehta 1993), of focus group outputs (Wilkinson 2004), and of facial gestures (Heath 2002), has now entered the mainstream methodology in social sciences.

The analysis of large volumes of data - including texts and time series – necessitates systems that can process these volumes in a timely manner, and that can be used to combine the results of these analyses. The ESRC e-Social Science **F**inancial **In**formation **Grid** (FINGRID) project demonstrated the use of Grid-enabled systems[1] for analysing large volumes of texts and time series, in combination, for studies of sentiment analysis in financial investing. FINGRID provided a novel integration of leading-edge analytical techniques for treatment of quantitative data (Monte Carlo simulation and wavelet analysis) with innovative information extraction techniques for analysis of textual data, used to 'detect' *market sentiment* in news reports[2]. Unlike analysis of captured and unchanging data undertaken in the majority of e-Science activities, FINGRID analyses *living data*: quantitative live streams of prices of financial instruments (c. 2GB) and a specific kind of qualitative data – news relating to the financial instruments (on average 5,000 news stories or 30MB/day). Such living data provides an additional challenge for the analysis: newly captured data may make results of the analysis irrelevant.

In this paper, we describe a method for automatically extracting *sentiment* from texts: the frequency of use of positive and negative sentiment words that may reflect a community's sentiment. We apply the method to a corpus of financial news text to: (a) automatically extract frequent key terms and symbols; and (b) automatically disambiguate key terms and symbols. Manual analysis suggests that spatial metaphors (*fall/rise*, *up/down*), speed metaphors (*accelerate* or *decelerate*) and biological metaphors (*growth/decay*) are prevalent in financial texts. The use of key words and metaphors in financial text appears to be governed by a set of rules that some authors refer to as a *local grammar* (Sinclair 1991, Gross 1993). Local grammars can be useful for disambiguation, filtering-out sentences that contain metaphorical words, and terms, used in a polysemous sense. These (disambiguated) sentiments can be time-ordered, by the time of arrival/publication of the news, and so a set of discrete values for sentiment can be created for a specific period of time – a time series of sentiment. This sentiment time series can be cross-correlated with other time series, such as share prices, to facilitate examinations into, for example, whether market sentiment is fuelled in any way by the written sentiments of the financial community. By way of conclusion, we describe how our method, developed for the analysis of financial texts and time series, could be extended and applied to studies in the sociology of crime and in the anthropology of ethnicity.

Motivation

Rationality, Bounded Rationality and Sentiment

A financial economist can analyse quantitative data using a large body of methods and techniques in statistical time series analysis on "fundamental data", related, for example, to fixed assets of an enterprise, and on "technical data", for example, share price movement (Reuters 1999). The economist can study the behaviour of a financial instrument, for example individual shares or currencies, or aggregated indices associated with stock exchanges, by looking at the changes in the value of the instrument at different time scales – ranging from minutes to decades. Meanwhile, financial investors/traders are trying to discover the *market sentiment*, looking for consensus in expectations, rising prices on falling volumes, and information/assistance from back-office analysts. It can be argued that sentiments have no room in a rational enterprise that involves transactions with quantities of

---

1 The Grid of FINGRID is reported in a companion paper (Gillam, Ahmad and Dear 2005).

2 Final Report and related papers at: http://www.computing.surrey.ac.uk/grid/fingrid/papers.html

money for financing business and for furthering international trade and commerce. The efficient market hypothesis suggests that quirks caused by sentiments can be rectified by the supposed inherent rationality of the majority of the players in the market (Samuelson 1965). Recent developments in financial economics, signified by the emergence of *derivatives* and *arbitrage*, show the triumph of rational reasoning: such instruments/strategies were created on the basis of mathematical models (Black and Scholes 1972), and the trading can be monitored using the self same models (Miller 1990). The assumption of overarching rational behaviour has been reviewed by Herbert Simon (1978/1992) and Daniel Kahnneman (2003), and arguments have been presented in favour of a model of *bounded rationality* where the actors in a given social situation prefer to ignore facts and trust their own version of reality and the efficient market mechanisms fail to operate (see, for example, Kindlberger 2001, Lakonishk, Lee & Poteshman 2003).

The analysis of the hopes and fears of actors in a marketplace is an area of strategic concern both academically and commercially: 'One possible definition of investor sentiment is the propensity to speculate. [..] sentiment drives the relative demand for speculative investments' (Baker and Wurgler 2004:5). In a departure from classical finance theory, but using purely quantitative data, Baker and Wurgler (2004) investigated the effect of 'sentiment driven mispricing' by investors on a cross-section of stock returns over a period of 1963-2001. Using six subjectively chosen 'proxies', the authors claim to have isolated the sentiment components of the proxies from 'business cycle components'. The authors have computed a sentiment index based on a multivariate analysis of the proxies, and found that the changes in the index are 'highly correlated and visibly line up with anecdotal accounts of past [stock market] bubbles'.

News Analysis and Sentiment Analysis

Qualitative research methods are being used in financial economics, and in sociological studies of financial markets, for systematically studying the hopes and fears of the traders, investors, and regulators in the analysis of the behaviour of the markets[3]. Questionnaire surveys are being used to compute investor/trader sentiment by Yale's International Centre of Finance, who publish investor and trader confidence indices monthly. The monthly indices show a systematic difference between attitudes of the individual investor and the institutional investor (Shiller 2003 and Yale 2004). The University of Michigan constructs and publishes a consumer confidence survey based on questionnaires (Michigan 2004). Mackenzie (2000, 2003) has used 'oral history' interviews, e-mails and telephone interviews, of key players involved in a failed arbitrage firm. The author examined the 'sociology of arbitrage' and demonstrated limits to rational analysis and the impact of sentiments in arbitrage trading.

Over the last decade or so, financial economists have attempted to quantify the impact of public information release, and private information arrival, on price fluctuations of financial instruments. DeGennaro and Shrieves (1997) have used the presence of pre-selected keywords in the headlines of news-wires as a measure of information related to currency trading. The numbers of headlines containing the keywords appear to vary with the fluctuations in the currencies being traded. Baestaens and van den Bergh (1995) had earlier reported similar results based on the analysis of headlines and market volatility.

Since 2000, the analysis of news wire has become selective and targeted. Some researchers choose news related to economic and financial topics – for example, news about employment

---

[3] For example, The *Journal of Behavioral Finance* (Lawrence Erlbaum Associates, Inc), was formerly the *Journal of Psychology and Financial Markets.*

– and distinguish between scheduled and non-scheduled news announcements; a study that focuses on 21 specific topics concludes that 'volatility [in the markets] increases in the pre-announcement period due to speculative or informed trades' (Bauwens, Omrane and Giot 2003). Other authors pre-select keywords that indicate change in the value of a financial instrument – including metaphorical terms like *above, below, up* and *down* – and use them to 'represent' news stories: a number of news stories are selected and the vectors, indicating the presence or absence of the movement keywords, are then automatically classified as to whether the news report is positive about a share price or is negative: Using this approach, Koppel and Shtrimberg (2004) have observed that a system can 'reliably' learn whether or not a news story will have a negative impact. Yet other authors use the frequency of collocational patterns: news stories for assigning a 'feel-good/bad' score to the story: 'Good' news stories appear to comprise collocates like *revenues rose*, *share rose*; 'Bad' news stories contain *profit warning, poor expectation*; 'neutral' stories contain collocates such as *announces product*, *alliance made*. The 'sentiment' of the story is then correlated with that of a financial instrument cited in the stories and inferences made (Seo, Giampapa and Sycara 2002).

A method for identifying and extracting sentiment

The news wire analysis outlined above depends either on the ability of the author to analyse the texts, and/or pre-selected keywords. In our own earlier work, dating back to 1996, we computed a sentiment time series by creating a list of movement keywords and then automatically analyzing financial news texts for the presence or absence of the keywords in the texts. The analysis leads to a positive <u>and</u> negative score for each news text - just the frequency of positive and negative keywords in the text. This score was aggregated over a specified time scale (hourly, daily, weekly, monthly, annually), and a time series of positive and negative 'sentiment' was obtained. The time series was correlated with the movement of a given financial instrument. In some cases, the time series of negative sentiment was highly correlated with a declining financial time series; at other times the time series of positive sentiment was highly correlated with a rising time series.

There are number of drawbacks to this rather *laissez-faire* approach to meaning in free natural language texts. First, free natural language texts are notoriously ambiguous: consider the tokens *rise* (and its variants *rose, rising*) and *fall* (and its variants *fell* and *falling*) that are used to describe an increase and decrease in the value of a financial instrument. A news report containing *rise/fall* in conjunction with an instrument may indicate a positive/negative sentiment. However, such conjunctive or *collocational* usage with, say *oil prices*, may have a converse effect for an economy at large, or for heavily oil-dependent sectors such as airlines, but perhaps is of benefit to the oil companies. Second, the words *rose/fall* are polysemous – rose may indicate an increase in value but could denote the flower, when capitalized may refer to a name, and *fall* may refer to the season as well. Third, the choice of a list of 'sentiment words' is subjective.

Ambiguous or not, news stories do contain indications of sentiment, and the volume of the stories is very large and growing – c. 5,000 stories per days comprising 30MB data per day on Reuters alone. Any analysis of such volumes requires a degree of automation for maximizing the potential of this growing and vital resource of streaming news. The ambiguity in such texts is bounded due to the 'special' language used in composing such stories. *Special languages* are written for a specific audience with the assumption that there is a set of shared concepts within a given domain of knowledge. These concepts are instantiated through a terminology (Sager, Dungworth and Macdonald 1988) and a *local*

grammar (Gross 1993). A special language is a subset of a natural language, for example an English special language of physics, or a Russian special language of rocket science. It is the frequent use of a select group of terms that distinguishes a special language from its mother - general language: the frequent use is idiosyncratic, can be used to signal a *term* and its collocates (Ahmad 1995), and can be used to provide a computationally tractable approach to identification of key terms (Gillam 2004). Essentially, the frequency of each word in a special language corpus is computed and contrasted with the frequency of the same word in a general language corpus. Tokens that occur with the same (relative) frequency in both corpora are deemed to be the basis of the natural language, while words that occur with much higher relative frequency in the special language corpus, or occur highly frequently in the special language corpus but not at all in the general language corpus, are designated as 'candidate' terms.

The candidates are analysed to find the strength of co-occurence with other candidates and, indeed, other words. Repetition of a pair is usually used to emphasise a certain attribute of one of the pairs; the use of *percentage* (symbol or token) is related to differentiation, as in *X% [percent] of shares of company Y were bought by company Z*, or to change in the value of an observable, for example *shares of company Y rose/fell [by] X% [percent]*. Collocative use of words, especially in specialist texts, are governed by the so-called *localist idiom priniciple* (Sinclair 1991): The principle suggests that although *rise/fall* belong to the (universal) category of verbs, these verbs can only be substituted by a small number of the members of the category when used in conjunction with *percent*. Collocations, Sinclair, argues are not formed by random chance, rather the restrictions placed on co-occurrence of words suggests that a *local,* rather than a *universal* grammar is used by the authors. This is particularly true for special language texts. The extraction of 'significant' collocates has generated substantial debate in the literature (see Armstrong 1994); one of the methods for extracting collocates we have followed is due to Smadja (1994). The re-collocation of an extant collocate, together with relevance feedback, leads to a set of rules (of a local grammar), expressible using a finite state automaton. The automaton can then be used to search an unseen text to identify patterns of discourse that suggest changes in the values and state of key objects and events in a specialist domain.

We adopt a text-driven and bottom-up method: starting from a collection of texts in a specialist domain, together with a representative general language corpus, and use the following five-step algorithm for identifying discourse patterns with more or less unique meanings, without any overt access to an external knowledge base:

    I.   Select training corpora: a randomly sampled special language corpus and a general language corpus.

    II.   Extract key words;

    III.   Extract key collocates;

    IV.   Extract local grammar using collocation and relevance feedback;

    V.   Assert the grammar as a finite state automaton.

We describe the use of this algorithm in the following section.

Experiments with and Evaluation of sentiment analysis method

Bootstrapping a local grammar

I. Training-Corpus Selection

There are a few representative general language corpora of the English language of everyday usage: The British National Corpus, comprising 100-million tokens distributed over 4124 texts (Aston and Burnard 1998), is a widely cited corpus and was chosen accordingly. The special language corpus chosen was the Reuters Corpus Volume 1 (*RCV1*) comprising news texts produced in 1996-1997 and contains 181 million words distributed over 806,791 texts. This corpus is available under licence from the US Nat. Inst. of Science and Technology.

II. Extraction of Keywords

The frequencies of individual words in the RCV1 were computed using System Quirk; for describing how our method works we will use a randomly selected component of the corpus – the output of February 1997, henceforth referred to as the RCV1-Feb97 corpus containing 14 Million words distributed 63,364 texts. Cumulative frequencies of the 50 most frequent words show that all the 50 words in the BNC, comprising 38% of the 100 million tokens, belong to the so-called closed class words (determiners, conjunctions, pronouns, modal verbs and prepositions). In the RCV1-Feb97 corpus there are 7 nouns and one verb (*said*) amongst the 50, again comprising 38 % percent of all the specialist corpus (see Table 1a):

Table 1a: Cumulative Frequency of Distribution of the 50 most frequent words in the corpora

Ranks	RCV1 Feb97 ($N_{RCV1Feb97}$=14 Million)	Cumulative Number of Tokens (%)	British National Corpus (N_{BNC}=100 Million)	Cumulative Number of Tokens (%)
1-10	the, to, of, in, a, and, **said**, on, s, for	0.87 M (21.3%)	the, of, and, a, in, to, for, is, as, that	22.3 M (22.3%)
11-20	at, that, was, is, it, by, with, from, **percent**, be	0.28 M (6.8%)	was, I, on, with, as, be, he, you, at, by	6.51 M (6.5 %)
21-30	as, he, **million**, **year**, its, will, but, has, would, were	0.17 M (4.2%)	are, this, have, but, not, from, had, his, they, or	4.23 M (4.2%)
31-40	an, not, are, have, which, had, up, n, new, **market**	0.13M (3.3%)	which, an, she, where, here, we, one, there, all, been	3.05 M (3.1%)
41-50	this, we, after, one, last, **company**, u, they, **bank**, **government**	0.10M (2.6%)	their, if, has, will, so, would, no, what, can, when	2.35 M (2.4%)

The relative frequencies of the closed class words are about the same in the two corpora, but there are systematic differences in the relative frequencies of the open-class words. We use a metric called *weirdness* that captures this difference, and has been successfully used to identify candidate terms (Ahmad and Rogers 2001). Table 1b presents the five most frequent candidate terms that have high weirdness values as well:

Table 1b: *weirdness* of the most frequent candidate terms in the RCV1 sample

Token	RCV1 Feb97 ($N_{RCV1Feb97}$= 14,244,349)			BNC (N_{BNC}=100,000,000)			Weirdness (a/b)
	Rank	$f_{RCV1Feb97}$	$f_{RCV1Feb97}/N_{RCV1Feb97}$ (a)	Rank	f_{BNC}	f_{BNC}/N_{BNC} (b)	
percent	19	65763	0.462%	3394	2928	0.003%	157.84
market	40	36349	0.255%	301	30078	0.030%	8.49
company	46	29058	0.204%	219	40118	0.040%	5.09
bank	49	28041	0.197%	562	17932	0.018%	10.99
shares	56	23352	0.164%	1285	8412	0.008%	19.51

III. Extract Candidate Collocates.

Collocation analysis is seeded using the frequently used candidate terms. We compute the frequency of contiguous (*percent rise* or *fall*) and non-contiguous collocations (*rose/fell by X percent*), over a window of five words to the left and right of a candidate. The frequency of occurrence of a collocate in each position is used to compute metrices for identifying statistically-significant collocates (see Smadja 1994). Collocates that exceed a threshold value of the metrices are deemed significant. Table 2 shows the use of *z*-scores (computed in turn from frequency information) in the identification of the collocates of *percent:*

Table 2. The frequency of *significant* 'left' and 'right' collocates of *percent* in the RCV1-Feb97. Details of the algorithm can be found in Smadja (1994), and that of a new implementation in Gillam (2004).

	f	Left	Right	Total	z-score
percent	65763				
up	5315	4360	955	5315	15.91
rose	4361	3988	373	4361	13.04
rise	2391	980	1411	2391	7.12
down	2291	1636	655	2291	6.82
fell	2074	1844	230	2074	6.17

IV. Local Grammar Extraction

The pattern *rose + percent* collocate has a smaller number of further collocates. Table 3 shows some of the collocates but with smaller z-scores.

Table 3. The re-collocation patterns of percent+rose.

Pattern	*f*	Collocate	Left	Right	z-score
10 percent to	108	*rose*	24	0	5.45
by 10 percent to	18	*rose*	5	0	2.27
rose 10 percent to	14	*billion*	0	7	4.24
rose 20 percent to	11	*billion*	1	7	6.02

There are other collocation patterns involving cardinal numerals (as in column 1 of Table 3; *rose five percent to, rose 4.5 percent to*) and many other collocates in addition to the ones involving *billion* (Column 3). These re-collocates and the display of keywords-in-context (*rise, fall, percent*) provide relevance feedback for the compilation of a local grammar.

V. Assertion of Local Grammar

The (re-) collocation patterns can then be asserted as a finite state automata for each of the movement verbs and spatial preposition metaphors (Figure 1 shows the automaton for *rose*):
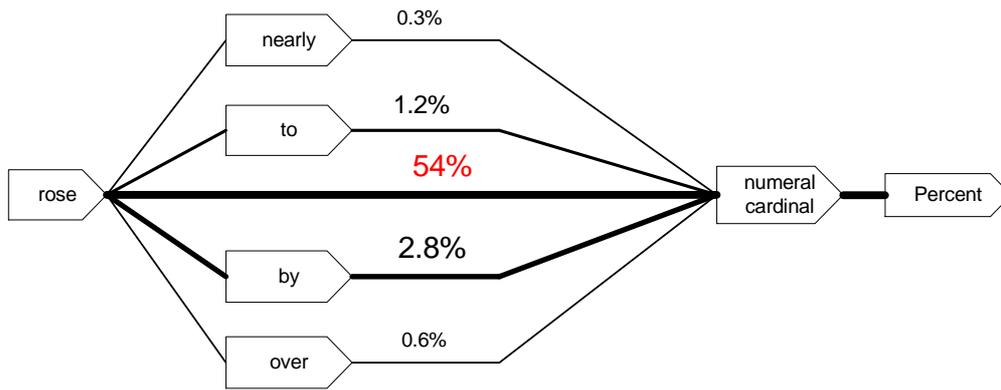
Figure 1: Collocation and re-collocation of *rose+percent* leads to the above automaton. The former accounts for over half the collocations – all contiguous - and the rest are accounted by other collocates. The percentage figures in the diagram show how many of the collocation patterns follow a given pattern; the largest being *rose+cardinal numeral+percent* (54%).

These local grammars can be to extract similar sentiment patterns from unseen, or newly arriving, texts. The results of this extraction will yield patterns that are, predominantly, associated with changes in the value of instruments.

Evaluatuion 1: Reducing the ambiguity

Use of the finite state automata for disambiguating sentiment extracted from financial news can be demonstrated by considering the frequency counts of positive and negative *tokens* found in unseen texts. Consider, for example, live news wire feed form Reuters Financial News Service for 15th-16th November 2004, comprising 6.64 Million words: There were 80,907 positive words and 43,636 negative words in the 6.64 Million. However, the number of these tokens used in the local grammar patterns extracted from the corpus analysis discussed above, is 1,927 positive and 3,924 negative words only. Sentences that 'obey' the local grammar are sentiment-bearing sentences. These sentences can be used as an input to a sentiment analysis; we may use the number of local-grammar obeying sentences as a proxy for positive/negative sentiment (Figure 2).
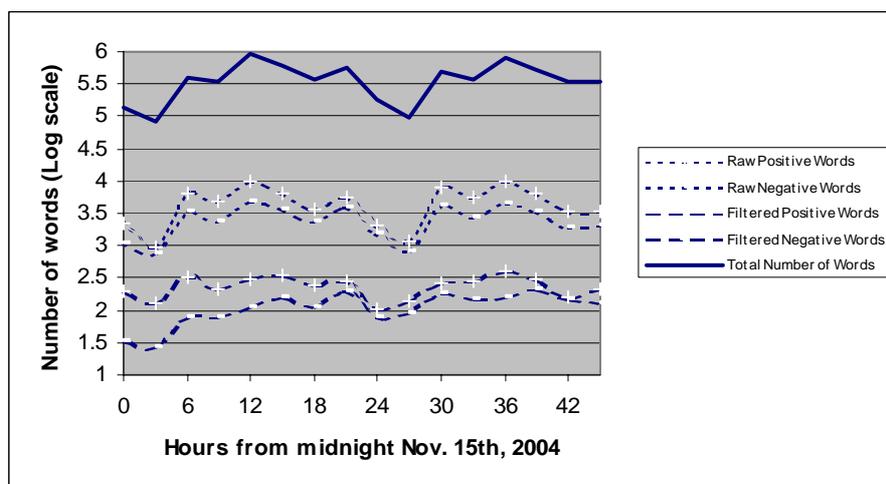

Figure 2. Changes in the total number of positive/negative words together with those that are used in the local grammars (filtered positive/negative words) and total number of words.

There is a need to investigate how the majority of the positive/negative words (over 95% of the total) are used outside the scope of the local grammar. There is a possibility that these words may not have been used to express change in values of financial instruments; terms like *percent* may be ellipsed; there may be a truly ambiguous use of the positive/negative words. Research on this topic is continuing at Surrey.

Evaluation 2: Increasing the throughput

Financial news is quite ephemeral in nature: the next item of news may contradict all that went before it; today's news should form the basis for tomorrow's patterns; sometimes there is a requirement to analyse news of up to 3, 5 or 10 years depending on the nature of the task, especially in the area of policy planning. The text corpus, for instance, may be analysed for the frequency distribution of words comprising the corpus. Consider the (complete) RCV1 corpus (181 million words in 806,791 texts): Processing this corpus as a single process on a single machine (a Dell PowerEdge 2650) in this configuration takes 53300 seconds.

To expedite such computations, we have created a 24 node grid infrastructure, which can provide access to upto 64 processors simultaneously, in an attempt to support such analyses (Gillam, Ahmad and Dear 2005). Using 16 processors we gain a throughput increase by a factor of 15 (3572 seconds); using 64 processors, the time is halved again (1683 seconds). Additional work is necessary to produce such parallel implementations: consideration has to be made of assigning analyses to different, available or already busy, processors; files may need to be uploaded to the processing machine or consideration made of fileserver bottlenecks; output results have to be merged, and so on.

Conclusions and Future Work

There are two principal drawbacks related to our method: first, though we have devised programs that can learn unambiguous patterns of use of positive or negative sentiment, a sentence is always used in the context of other sentences and the context may change if the inference is made on the basis of one sentence only; second, one can argue that a new text is a response to some or all of the existing texts, and in that sense each text is contextualised within a network of other texts - even if all the existing texts unambiguously expressed a positive sentiment, a new text with strong negative sentiment may invalidate all of the positive sentiment.

The range of quantitative analysis techniques tested, implemented and evaluated in FINGRID, and in developments beyond FINGRID (but not reported in this paper), includes wavelet analysis (Ahmad et al 2004), fuzzy-logic knowledge bases (Poopola et al 2004), and case-based reasoning. The results of the two analyses may be used to create a *confidence index* – or *sentiment index*. The techniques developed in the FINGRID project for conjunctive textual and quantitative analysis, can, in principle be extended to the new areas identified in the introduction. In a recent proposal, we presented such an extension by specifying different text types that may be of interest to the three social science areas: *informative* – including news reports, *appelative* – commentaries on financial performance of enterprises and governments, and 'Letters to the Editor' sent by citizens describing their perceptions of crime and ethnicity, and *expressive* texts obtained using semi-structured interviews and focus group outputs. These texts can be automatically analysed using methods of corpus linguistics, discourse analysis and information extraction (IE). Quantitative analysis methods developed in the FINGRID project can be used in the analysis on-line or accessible data such as crime statistics, for sociology of crime, and labour force

surveys, based on race/ethnicity for anthropology. The fusion of the results of the textual and quantitative analysis can, in turn, be used to automatically produce a *crime confidence index*, for measuring the fear of crime (Fielding, Innes and Fielding 2002), and a *conflict index*, for measuring ethnic/racial tension in a community (Eade and Samad 2002). We present a matrix that relates the quantitative and qualitative data that can be analysed using appropriate methods and techniques described above. The framework describes our vision of e-mediated economics, sociology and anthropology.

Table 4. An integrated view of using current and emerging analytical techniques for analysing large-scale and distributed qualitative and quantitative data sets in e-Social Sciences

	iNVESTOR PSYCHOLOGY	SOCIOLOGY OF CRIME	ANTHROPOLOGY OF ETHNICITY	Methods/ Techniques
Qualitative data INFORMATIVE				
	Financial News and Reports; State-of-the-Economy Reports; Company Reports.	National News Reportage & Editorials; Police Authority & Other Reports; Policy Documents	National and International News Reportage & Editorials; Local Govt. Reports; Policy Documents	CORPUS LING. & IE: Terminological, Grammatical and Ontological Analysis for Identifying and Disambiguating sentiment and named entities
APPELLATIVE	News Commentaries on financial instruments.	'Letters to the Editor'; Web Sites	'Letters to the Editor'; Web Sites	DITTO
	Focus Group Encounters	Semi-structured interviews	Semi-structured interviews	DISCOURSE ANALYSIS
EXPRESSIVE	Executive movements; corporate entity identification		Anonymisation of field data	IE: Named Entity extractors
Quantitative data HIGH FREQUENCY (NUMERICAL)				
	Technical Data (e.g. Stock Price Movement; Price/Earning Ratio)	Crime Statistics	Labour Force Surveys; Educational Achievement Surveys	WAVELET ANALYSIS; MONTE-CARLO TYPE BOOTSTRAPPING
LOW FREQUENCY (NUMERICAL)	Company demographics – fixed assets	UK census data	UK census data	DATA ANALYSIS; AGGREGATION; VISUALISATION; CASE-BASED REASONING (CBR)
INDETERMINATE	Questionnaires	Questionnaires	Questionnaires	DITTO
Fusion	Confidence Index	Crime Index	Conflict Index	DATA MINING; VISUALISATION TECHNIQUES
	Investment decision (buy/sell)	Policy formation / evaluation	Policy formation / evaluation	ONTOLOGY LEARNING FOR RULE-BASED / CASE-BASED REASONING

Acknowledgments

References

Ahmad, K. (1995). Pragmatics of Specialist Terms and Terminology Management. In (Ed.) Petra Steffens. *Machine Translation and the Lexicon.* (LNAI, Vol. 898) Heidelberg: Springer. pp.51-76

Ahmad, K., and Rogers, M. A. (2001). "Corpus Linguistics and Terminology Extraction". In (Eds.) Sue-Ellen Wright and Gerhard Budin. Handbook of Terminology Management (Volume 2). Amsterdam & Philadelphia: John Benjamins Publishing Company: 725-760.

Ahmad, S., Taskaya Temizel, T., and Ahmad, K. "Summarizing Time Series: Learning Patterns in 'Volatile' Series." Z.R. Yang, R. Everson, and H. Yin (Eds.). *Proc. of 5th Int. Conf. on Intelligent Data Eng. and Automated Learning (LNCS Vol. 3177).* Heidelberg: Springer Verlag. pp 523-532.

Armstrong, S. (1994). (Ed.) *Using Large Corpora.* Cambridge (Mass.) & London: The MIT Press.

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

Baestaens, D.J.E. and van den Bergh, W.M. (1995). 'The marginal contribution of news to the DEM/USD swap rate.' *Neural Network World*, Vol. 5(No. 4), pp. 371-378.

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

Black, F., & Scholes, M. (1973). 'The Pricing of Options and Corporate Liabilities'. *Journal of Political Economy*, Vol 81, pp 637-654.

Dalgleish, D., & Myhill, A. (2004). *Reassuring the public: a review of international policing. (Findings 241 - Home Office Research Study No. 284.)* London: Home Office (http://www.homeoffice.gov.uk/rds/pdfs04/r241.pdf)

DeGennaro, R., and R. Shrieves (1997): 'Public information releases, private informationarrival and volatility in the foreign exchange market". *Journal of Empirical Finance* Vol 4, pp 295–315.

Eade, J,. and Samad, Y. (2002) Community Perceptions of Forced Marriage. London: Foreign and Commonwealth Office. (available at http://www.fco.gov.uk/Files/kfile/clureport.pdf).

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. (2004). *Systems of Concept and their extraction from text.* (PhD Dissertation). Guildford: University of Surrey.

Heath, C. C. (2002). 'Demonstrative Suffering: the gestural (re-) embodiment of symptoms. *Journal of Communication*. Vol. 52(No.3). 'pp 591-617.

Jackson, R. (2004). 'The Social Construction of Internal War' In (Ed.) Richard Jackson. *(Re)Constructing Cultures of Violence and Peace*. Rodopi: Amsterdam/New York.

Kahneman, D. (2002). Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice (A Nobel Prize Lecture December 8, 2002). (Available on http://nobelprize.org/economics/laureates/2002/kahnemann-lecture.pdf).

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

Kindleberger, C. P. (2001). *Manias, Panics, and Crashes: A History of Financial Crises* (Wiley Investment Classics). New York. John Wiley & Sons.

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

Lakonishok, J., Lee, I., and Poteshman, A.M. (2004). "Investor Behavior in the Option Market" (January 2004). NBER Working Paper No. W10264. Cambridge: National Bureau of Economic Research. (Available at http://ssrn.com/abstract=495769)

Mackenzie, D. (2000). 'Fear in the Markets'. *London Rev. of Books.* Vol 22 (No. 8).

Mackenzie, D. (2003). '*Long-Term Capital Management* and the sociology of arbitrage'. *Economy and Society* Vol. 32 (No. 3). pp 349-380.

Markham, A. N. (2004). Internet communication as a tool for qualitative research. In (Ed). David Silverman. pp 95-124.

Metha, J. (1993). Meanings in the context of bargaining games: Narratives in opposition. In (Eds.) W. Henderson, Tony Dudley-Evans and Roger Backhouse. pp 85-99.

Michigan (2004). Survey of Consumers - May 2004 (published June 2004). (Available at http://www.sca.isr.umich.edu/)

Miller, M. H. (1990). 'Leverage'. In (Ed.) Karl-Göran Mäler. Nobel Lectures in Economic Sciences 1981-1990. Singapore: World Scientific Publishing Company. pp 291-300 (Also available on http://nobelprize.org/economics/laureates/1990/miller-lecture.pdf)

Popoola, A., Ahmad, S., and Ahmad, K. "A Fuzzy-Wavelet Method for Analyzing Non-Stationary Time Series." *Proc. of The 5th Int. Conf. on Recent Advances in Soft Computing (December 16-18, 2004, Nottingham, UK).* (http://www.computing.surrey.ac.uk/grid/fingrid/papers.html)

Samuelson, P. A. (1965). 'Proof that Properly Anticipated Prices Fluctuate Randomly'. Industrial Management Review, Volume 6, pp 41-49.

Shiller R. J. (2003). The New Financial Order: Risk in the 21st Century. Princeton: Princeton University Press.

Silverman, D. (2004). (Ed.) *Qualitative Research: Theory, Methods and Practice.* London: Sage Publications.

Simon, H. (1992). 'Rational Decision Making in Business Organsations (A Nobel Memorial Lecture, 8 December, 1978).' In (Ed.) Assar Lindbeck. Nobel Lectures in Economic Sciences 1969-1980. Singapore: World Scientific Publishing Company. (Also available on http://nobelprize.org/economics/laureates/1978/simon-lecture.pdf)

Smadja, F. (1994). 'Retrieving Collocations from Text:Xtract'. In (Ed.) S. Armstrong. pp 141-177.

Wilkinson, S. (2004). 'Focus Group Research'. In (Ed.) David Silverman. pp 177-199.

Yale (2004). Yale School of Management Stock Market Confidence Indexes™. Available at http://icf.som.yale.edu/financial_data/confidence_index/index.shtml