Edderkoppspinn eller nettverk: News media and the use of polar words in emotive contexts

Khurshid Ahmad

Department of Computer Science

Trinity College

Dublin

Summary
An adaptation of two well-established measures of changes in financial markets – *return* and *volatility* – is presented for the analysis of changes in sentiment, articulated in text, towards specific groups of people or towards an identifiable system of beliefs. The method used in this analysis of sentiments is based on the use of a thesaurus of affect words and on the computation of the rate of change of these words over time in a diachronically organized corpus of texts. A corpus based analysis of news about comprising *Islam* and *Muslim* as keywords is presented covering three important periods – calendar years 1991, 1999 and 2007 (circa 700,000 words) from one of the most prestigious US-based newspapers, *The New York Times*.

Introduction

Sentiments and Connotation

The choice of words in a text is usually a deliberate choice on part of the author within his or her institutional, social, and professional context. The scientist-author writes to express his or her own *ontological commitment,* a commitment to a theory or to an experimental method that allows him or her to seek knowledge about what there is (ontology can be defined as a study of *what there is,* see Ahmad 2007a for details). The ideologue-author writes to express his or her own commitment to a system of thought and emotion and attitude to the world, to society and to humanity at large. One can argue that the semantics of ontological commitment and ideological commitment are not very different: scientists when they reject or affirm an ideology using their own ontological commitment are criss-crossing between the so-called rational scientific method and 'fundamentalism' of sorts. Similarly, the ideologues who seek sustenance from the ontological commitments found in natural, physical, biological, engineering and other sciences.

The ontological and ideological commitment manifests itself through the repetitious use of certain keywords and the collocations of the keywords. In physics, we have *energy* and *force* and the collocates of these two keywords; in modern biology *evolution, gene* and *cell* are important pennants used by many biologists and a number of collocates of these keywords are used to populate biological literature. The same is true, perhaps to a lesser extent of human sciences, where the choice of terms appears less regulated than is the case in physical, biological and engineering sciences. But pennants fly here as well: *grammar* was a byword for Chomskyian linguistics and it had many 'surface' and 'deep' meanings. Keywords used by scientists are sometimes metaphorical in nature and extant words are assigned relatively new meanings and the meaning propagated through repeated usage. The case of the term

nucleus is an interesting one: starting from the 'centre' of celestial universe, i.e. the Sun, the physical scientists used this metaphor of the centre to predicate the existence of the 'nucleus' of an atom and the biological scientists found a similar object at the centre of a cell.

During times of conflict the choice of words becomes critical in the ideologically motivated literature. The near-synonyms like *freedom* and *liberty* were used during the Cold War in Europe and North America to indicate whether the author was 'right leaning, that is those using 'freedom', or 'left'-leaning, i.e. those using 'liberty': there was an implication of a word or phrase in addition to its literal meaning – the connotation of a word. It has been argued that sentiments can be expressed by using certain particular words to articulate connotative meaning.

The use of connotation is quite popular for describing new or different groups people, places and things. Connotation can also be used to identify, isolate and celebrate or denigrate people, places and things. This is a big project and I have made it more difficult for myself by exploring the hypothesis that we can build a machine that will be able to identify and elaborate sentiment-laden phrases, clauses and indeed whole stretches of texts. Foolhardy perhaps but exciting and adventurous nevertheless.

A personal note

Some 15 years ago I was trying to understand the coinage, usage and obsolescence of specialist terms by examining the distribution of words in special language corpora. A method that I had developed, with the co-operation of my doctoral students, research assistants and a number of colleagues (Ahmad and Rogers 1992 & 2001, Kugler, Ahmad and Thurmair 1995), appeared to work well for texts written in English; we would look at the distribution of words in a specialist corpus of texts and then look at the frequency distribution of the same words in a representative corpus of English used for general purposes. The basis of the method was very simple: the specialist terms will have a higher (relative) frequency of occurrence in a specialist corpus than will be the case for a general language corpus. The ratio of the relative frequency of the same words in the two corpora will help in making an objective judgment about the termhood of the term. The ratio I had called *weirdness* – a remark Bronislaw Malinowski made about the language of South Sea Island shamans in that there was profusion of nouns in the language of shamans when compared with the average islanders. Having been trained as a scientist, I reflected a bit and thought that this is what scientists do when they write – they use a language full of nouns sometimes almost impenetrable to the wider public. The language used by scientists inculcates the same awe in the minds of the general public in the proximity of the scientists as do the shamans in the minds of the average islanders in the South Sea.

The term *weirdness* and my proselytization of corpora raised an eyebrow or two in the hallowed environments of normative, Wusterian terminology. Two of its key Norwegian proponents, Johan Myking and Magnar Brekke become friends despite their raised eyebrows. I did persuade them to co-author a paper on Norwegian specialist terminology on the arachnology – we worked on a set of articles in Norwegian describing spiders. This was 1996 and there was not much in the artificial spiders web (the World Wide Web); nevertheless, a corpus was constructed. We found that we could extract a set of terms related to spiders in Norwegian following the corpus based methods I have alluded to above (Brekke, Myking and Ahmad 1996).

More recently, I have attempted to look at how language is used in situations where there is uncertainty about the value of objects of desire – for instance, price of a stock or share in a financial market. Here the connotative meaning plays a key role in expressing a feeling about the value of stock or share of a company, for example, which may either be in a booming state or is about to go bust. In such turbulent situations, connotative use of language may be linked to an attempt by the stakeholders in the market to discover the actual value of the share. It has been argued that the sentiment of the investor or the trader is influenced by what the investor reads or hears; *news impact analysis* is a topic that is hotly debated amongst economists and finance scholars – some believe that all the information about any economic entity is in the monetary value of the entity and the positive/negative effects of the news are discounted; whilst others believe that news has an impact and negative news has a larger impact than the positive news. Sentiment analysis uses the terminology of a specialist domain, say the terminology associated with anyone of the financial instruments (e.g. stocks and shares, currencies, bonds, market indices) in conjunction with a thesaurus of words that are used in articulating sentiment. The sentiment count is offered as an index by major financial news vendors like Reuters (for a brief overview, see, for example, Ahmad 2008).

This paper takes the study of connotative meaning in a more turbulent area of expression of feeling about a group of people – groups categorized on religion, ethnicity, culture, entertainment and so on. I have attempted to introduce a frequency based study of the use of keywords for studying changes in 'sentiment' related to a group over time. This study is based on a selection of texts in a newspaper over a quarter of a century. I hope, a study of diachronic change in frequency of (key)word usage, together with the contiguous affect words, may help to identify the sentiments of a people about the Other.

The study of the distribution and profusion of affect words in speech or text by somebody speaking or writing with an intention of persuading others, is becoming fashionable in economics and finance (see, Tetlock 2007 for example). However, this profusion can be found in one of the oldest genres, *religious writing*, and affect words are used, sometimes with abandon, by and about people with religious beliefs. The confluence of the two themes, economics and religion, appears very apposite when I think of Magnar Brekke. Over the last 20 years or so, I have had long and enjoyable discussions with Magnar on two topics: First, terminology, especially corpus-based terminology, and the language of economics and finance; second, on a personal level, discussions about religious beliefs. This article is being offered as a celebration of his life-long dedication to improving communications amongst people of different kinds and his abiding interest in all matters linguistic.

To sum up then: This paper is an attempt to understand how connotative meaning is communicated in texts. I have attempted to use measures of turbulence, developed in economics and finance for looking at the dynamics of price changes in a market, for examining changes in the distribution of *affect* words. I have looked at the diachronic distribution of affect words in news items that appear to cover topics related to *Muslim* and *Islam*. These news items were published in a representative sample of American English, *The New York Times*; the items that were examined can be viewed on the *New York Times* online archive. The diachronic study focused on the monthly frequencies of positive and negative affect words, the rate of change of the frequencies, and variance of the rate of change. In economics and finance, the rate of change of price of an asset is sometimes called *return* and the standard deviation of *returns* over a period of time is called the *volatility*. The return and volatility of affect words is presented in this paper as a measure of changes in sentiment, articulated through the use of affect words, about a group of people or a system of beliefs.

The Data: Searching the *New York Times* Online Archive

The selection of texts for a corpus based analysis is always a contentious subject and the corpus-builder's bias. Now the choice of texts for building a corpus to study language is in itself a rather controversial topic (see, for example, Ahmad 2007b). I have made my life even more difficult by choosing to study a much debated subject − religion. In mitigation, I have followed some of the tenets of corpus linguistics – I chose texts of different genres written by a number of different authors at different times. I have chosen newspaper texts for my analysis: the text types in a newspaper typically comprise reportage, opinions and editorials, travelogues, book and film reviews, and letters-to-the-editor, and not forgetting the advertisements.

I have chosen *The New York Times* (NYT), a prestigious US newspaper known for its independence and integrity. It has an on-line version that also allows access to its archive that spans over 150 years. This is a veritable source of information and opinion and a record of happenings within the USA and much beyond. The NYT, for me, is a microcosm of American English writing, where a balance between different political/social tendencies is maintained by employing opinion writers across the political/social spectrum.

I have chosen to look at diachronic change in sentiment of the contributors to the NYT to the religion of Islam and its practitioners, the Muslims. I have only chosen those texts that have at least one occurrence of the keyword *Islam* and one of *Muslim.* The diachronic change in sentiment can be studied just by looking at the *rate of change* of the frequency of usage of the keywords in the first instance over time. Latterly, one can study changes in sentiment by looking at the distribution of words usually used to articulate connotative meaning – for example 'valence' words that express positive or negative sentiment. The NYT search engine also retrieves texts that may comprise derivations like *Islamic* and *Islamization* or affixes like *Muslims.*

For the diachronic study, I first sampled texts from one period and computed the frequency of usage of the two keywords and the frequency of usage of sentiment words. Then another period of time was chosen, frequencies computed and then compared with the previous period. Three periods were chosen: (i) 1991 – the end of the First Gulf War, (ii) 1999 – a period of relative calm with the creation of a Palestinian state and rapprochement between Israel and the Arabs; and (iii) 2007 – the aftermath of the Second Gulf War and the onset of recession in the West.

The total number of such stories carried in the NYT has increased from 124 in 1991 (comprising 140326 tokens) to 155 in 1999 (192467 tokens) to 327 in 2007 (413939 tokens). The concomitant average rising from 12 items per month in 1991 to 13 in 1999 to double the figure of 27 in 2007 (see Table 1).

Table 1. Stories that appeared in the NYT with the keywords *Islam* **and** *Muslim* with samples drawn over three 8 year intervals, 1991, 1999 and 2007

Year	**1991**		**1999**		**2007**	
Month	**Items**	**Tokens**	**Items**	**Tokens**	**Items**	**Tokens**
January	15	1614	8	888	33	1645
February	13	1058	18	1209	23	1239
March	21	772	16	943	31	1083
April	7	1075	11	1313	39	1229
May	7	1072	9	1139	37	1211
June	12	1672	15	1402	35	1191
July	7	985	10	993	31	1388
August	4	1209	12	1147	19	1011
September	12	1342	10	1271	21	1236
October	9	796	19	1529	19	1183
November	5	718	13	1126	19	1461
December	12	787	14	1619	20	1238
Total	***124***	***140326***	***155***	***192467***	***327***	***413939***
Average	***10***	***1115***	***13***	***1215***	***27***	***1260***
Std Dev	***5***	***280***	***4***	***225***	***8***	***169***

The increase may also be attributed to the fact when news items are printed on paper there is an inherent restriction on length and the number of items. Electronic publishing allows one to overcome such restrictions. The doubling of the number of items relating to *Islam* may be attributed to this electronically-facilitated mode of communication. So, I carried out an ad hoc experiment on three corpora of texts that comprises stories published during 1981 to 2007 about the followers of one of the three religions with major following in the USA – *Christian, Jew* (or *Jewish*), or *Muslim.* The period covers texts that were only printed on paper (circa 1981-1991) and a period when it was the norm to have electronic editions (circa 1997-2007), and periods in between. My approximate computations tell me that stories containing *Muslim* had a minimum in 1983 (22 stories) and a maximum in 2001 (1754), stories comprising *Christians* had a minimum in 1993 (1280 stories) and a maximum in 2005 (2294 stories), and stories containing *Jews* (or *Jewish*) had minimum in 1983 (1733 stories in all) and a maximum in 2000 (2518 stories) (see Figure 1a). A similar experiment was conducted on three other corpora, one containing stories comprising at least one occurrence of *Christianity*, the second containing stories about *Islam* and the third about Judaism[1] (see Figure 1b). It appears that the electronic editions have contributed to an increase in the number and length of the stories, but variations in the verbiage has its own idiosyncrasies independent of the increased number and length. And, the interest in *Muslims*, reflected by the number of stories on this topic in the *NYT*, did catch up with the interest in the followers of other religions; the interest in *Islam*, always higher than in other religions, has really boomed after 2000.

Note that the six sub-corpora were built using a more relaxed criterion that any story that will be incorporated in the sub-corpora will have at least one occurrence of a keyword (*Christian,*

[1] Islam – minimum in 1983 (115 stories) and maximum in 2003 (1264), Christianity – minimum in 1983 (154) and maximum in 2004 (590), and Judaism – minimum in 1982 (82) and maximum in 1997 (236 stories).

Jew, or *Muslim*) – we will get lesser number of stories if we used the criterion that the stories contain at least one occurrence each of two terms (*Muslim* and *Islam*).

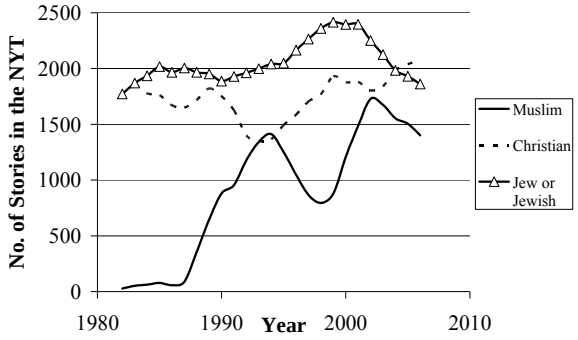

Figure 1a

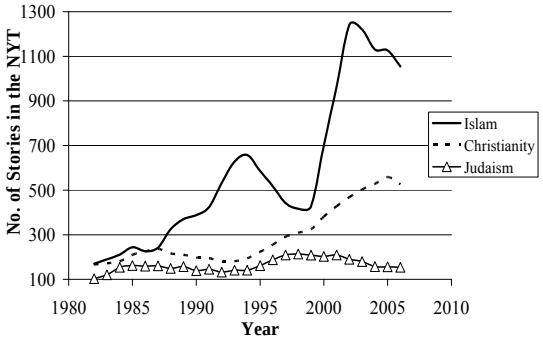

Figure 1b

Figure 1. Changes in the number of stories that had one or more occurrences of either names of the followers of the religion (Figure 1a) or names of religions (Figure 1b). The variation in the number of stories comprising the token *Muslim* is large –an average of 887 stories per year with a standard deviation (SD) of 617, whereas the number of stories comprising one or more occurrence of *Christian* have an average of 1744 (SD= 231). The average for stories containing *Jew* or *Jewish* is 2041 (SD= 210) (Figure 1a). The variance in the name of religions follows an inverse pattern (The average for Islam is 582 (SD= 379); for Christianity it is 298 average (SD=139) and for Judaism the average is 161 (SD =37); Fig 1b).

The rapidity of the increase in the number of stories is countered somehow by the changes in the number of stories for one year to the next – there is an oscillation in the number of items between 1985 onwards and one can discern more peaks and troughs in the number of items carried in the newspaper relating to Islam as compared to the other two religions (see Figures 1a and 1b). This brings me to the twinned notions of *return* and *volatility*. These terms are used extensively in the econometric and finance literature and are used to compute the rate of change in the value of various financial instruments. Essentially, *return* is described as the logarithm of the value of an instrument measured on two successive time intervals – hours, days, weeks, months and so on. So, if the number of news items containing the term *Muslim* in the NYT in the period 2004, 2005, and 2006 are 1189, 940 and 1251 respectively then the corresponding *returns* for 2005-2004 and 2006-2005 are:

2005-2004	- 0.102	(=log(940/1189))
2006-2005	+0.124	(=log(1251/940))

The considerable variance in the number of stories published containing the term *Islam*, and the lesser variance in the number containing *Christianity* and *Judaism* is reflected in the rapidity of change in the *return* values. The values are the logarithmic ratios of the number of stories published in two successive years from 1981 to 2005 (see Figure 1c).

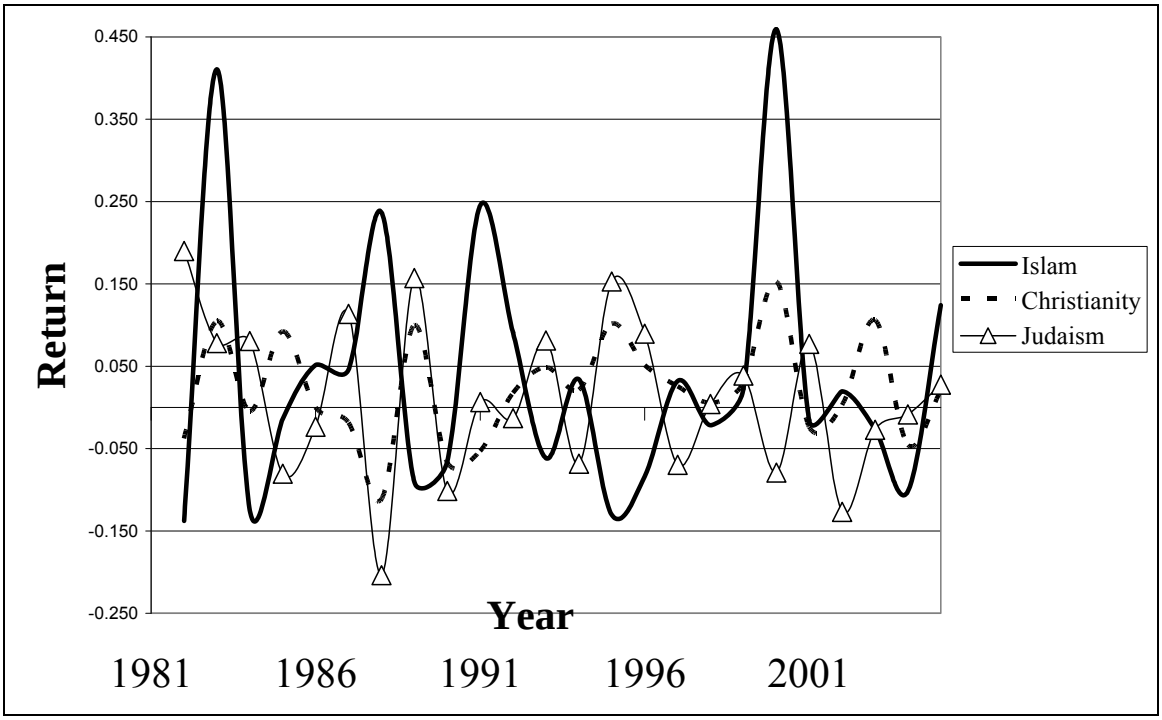

Figure 1c: Returns associated with the number of stories for two successive years in each of the three corpora containing (*Islam, Christianity* and *Judaism*) in each of the three corpora from 1981 to 2005

How do we quantify the rapid or not so rapid change in the number of stories, or *return*, about a topic? The key here is to use the *volatility of return*: the standard deviation of return. To illustrate the use of volatility, I have looked at the number of stories published at the beginning and published at the end of a 5 year period for a corpus of texts − where each text has at least one occurrence of the word *Islam.* This computation will result in the volatility related to the number of stories comprising the keyword *Islam* in that 5 year period: I have computed the volatility of these returns for a 24 year period from 1981 to 2005. Similar computations were carried out for two other corpora comprising texts published in same period – one corpus has texts where each text has at least one occurrence of *Christianity* and the other corpus comprises texts where each text has at least one occurrence of *Judaism* (see Fig 1d).

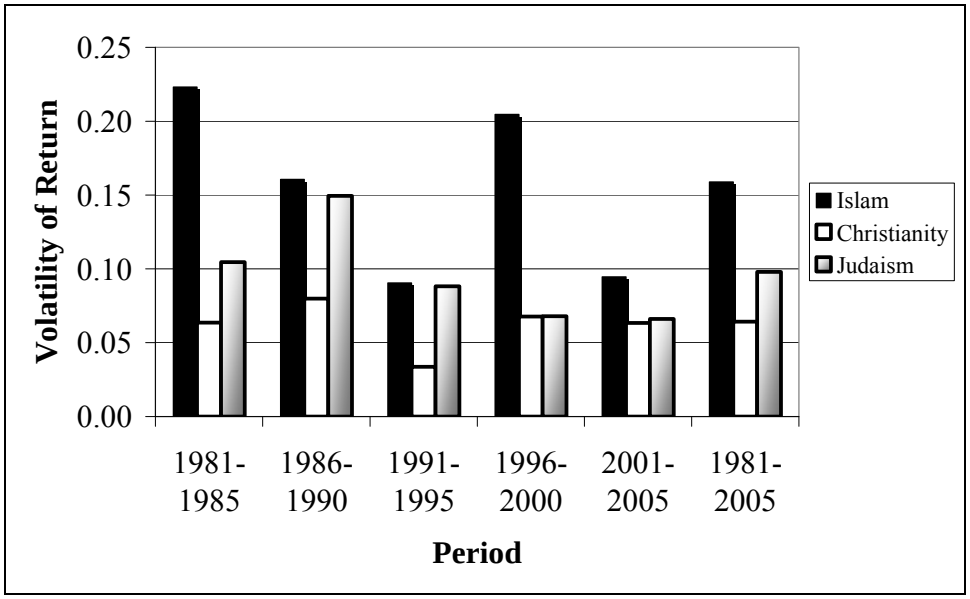

Figure 1d: Volatility associated with the changing frequency of the three terms , *Islam, Christianity* and *Judaism*, in each of the three corpora, comprising stories from the *NYT,* over 5 year intervals between 1981 and 2005

The reader will note that the volatility of returns is more nuanced for the corpus of texts on *Islam.* The volatility of returns associated with the other two corpora is less nuanced for texts comprising the word *Christianity*, we note very little change over the years and the standard deviation of returns is around 0.06. The volatility associated with texts comprising *Judaism* varies around 0.1, with a peak in the 1986-1990 period of 0.15. However, volatility of returns related to the number of stories published about *Islam* varies around 0.16 over the 24 year period. We now briefly look at how to interpret changes in volatility.

It has been argued in the literature on economics that "[a]s time goes by, we get more information on these future events and re-value the asset. So at a basic level, financial price volatility is due to the arrival of new information. Volatility clustering is simply clustering of information arrivals. The fact that this is common to so many assets is simply a statement that news is typically clustered in time." (Engle 1993:330). One can then argue that the larger volatility noted in the number of stories covered during a fixed period of time perhaps indicates that the writers of the stories are trying to discover more about the key topic covered in the stories.

The number of stories about a particular topic is a measure of interest in or curiosity about the topic in most general terms. What about the sentiments expressed related to the topic itself? Again, a long term project, a foolhardy enterprise, but interesting nevertheless as it leads to the question: Can we attempt to automatically interpret the contents of texts?

Well, there are serious attempts to understand 'value change' in a specific segment of human society by understanding the written language used by the segment, and *contents analysis* is one such subject. *Contents Analysis* is a name used to describe a collection of computer-based methods and techniques used to study the frequency distribution of a set of pre-specified (single-)terms drawn from a thesaurus – each term has one or more pre-assigned categories. *Contents Analysis* is popular in sociology, politics and increasingly in economics

where the distribution of terms within different categories is used to describe changing political, economic and social values, processes and systems. For example, in Namenwirth and Lasswell 1970, a pioneering study in *Contents Analysis*, the authors have looked at words that may relate to *power, rectitude, respect, affection, wealth, well-being, enlightenment* and *skill,* to look at the changes in the early (1844-1864) and late (1944-1964) 'platforms' or manifestos of the Democrats and Republicans to compare the differences within and across the two parties. The result of the analysis showed that as time went on, the manifestos were written using words of the same category with the frequency of the words in the two manifestos becoming closer as the time passed. This can lead us to the observation that since the manifestos were becoming very similar in their (connotative) language, then the political differences between the two parties were disappearing as well! *Contents Analysis* is being used to study the distribution of polarity-indicating words in order to ascertain the intentions of the writer with respect to other peoples, places and things. A tall order but an interesting adventure in text analysis which is sometimes referred to as *sentiment analysis*! This I will outline next.

Sentiment Analysis

Sentiment is defined as "an opinion or view as to what is right or agreeable" and political scientists and economists have used this word as a technical term. When sentiments are expressed through the faculty of language, we tend to use certain literal and metaphorical words to convey what we believe to be right or agreeable. There are a number of learned papers and reviews in computational sentiment analysis that are available (see, for example, Kennedy and Inkpen 2006 and references therein).

Harvard Dictionary of Affect

One of the pioneers of political theory and communications in the early 20th century, Harold Lasswell (1948), has used sentiment to convey the idea of an attitude permeated by feeling rather than the undirected feeling itself. This approach to analysing contents of political and economic documents – called content analysis – was given considerable fillip in the 1950's and 1960's by Philip Stone of Harvard University who created the so-called General Inquirer System (Stone et al. 1996) and a large digitised dictionary – the *GI Dictionary* also known as the *Harvard Dictionary of Affect*.

Stone et al., inspired also by Laswell et al., created the *Harvard Dictionary* which currently comprises over 11,000 words. Each word in the Dictionary has one or more 'tags'. Some of these tags refer to the connotative meaning of the word, whilst others to its cognitive orientation, and some to the belongingness of the word to a specific domain. The words in the Dictionary have between one and 12 of the 128 'tags'. These tags are divided into 28 or so categories. The *Harvard Dictionary of Affect* is based closely on Charles Osgood's attempts at the quantification of connotative meaning of words through the postulation of a semantic space. The space is a three dimensional space bounded by three primary factors or axes– evaluation (positive/negative), potency (strong/weak) and activity (active/passive). Osgood used Roget's Thesaurus to create bipolar scales, essentially a set of adjectives and their antonyms like *good/bad*, *pleasure/pain*, *vice/virtue* and so on. Eight categories were populated by including a wide range of bipolar words, especially around 2,000 or so words each in the evaluation category of positive and negative connotation words that Stone et al. call *valence words* (Table 2a):

Table 2a: Some of the words that are used to express 'evaluation', many are verbs together with a few nouns

'Valence' Category	**Verb-like**	**Noun-like**
Positive	aid, assist, associate, attract, collaborate, contribute, co-operate, defend, endorse, enthuse, inspire, join, offer, protect, shelter, spare, support, trust	passionate, trustworthy
Negative	abolish, abuse, aggravate, antagonize, arrest, banish, bar, beat, bomb, capture, compel, combat, constrain, convict, exclude, exploit, fine, force, hamper, impair, jail, knock, limit, manipulate, murder, neutralize, oppose, overpower, rebel, repulse, seize, shock, threat, thwart, withhold	aggressive, force, guerrilla, merciless, monster, pandemonium, repudiate, violence, vengeance

There are seven more categories of bi-polar words in addition to the valence *positive/negative* words including two of Osgood's other categories of *potency* and *activity* (the Osgoodian *evaluation* category is subsumed in the valence category). In four categories there are four times as many positive polarity words as there are negative polarity words (these categories are labeled motivational adjectives, potency, activity, and emotional expressiveness); in three categories there is an equal distribution (evaluation, pleasure/pain, valence). Only in one, rather obviously named category of *negation*, the ratio of negative to positive polarity is 10:1. There are a few 'neutral' words in three other categories (see Table 2b).

Table 2b: The distribution of *bi-polar* words in *Harvard Dictionary of Affect*

Categories	**Polarity**				**Other**	
	+	*f*	-	*f*		*f*
Valence	Positive	1915	Negative	2291		
Osgood Sem. Diff: Activity	Active	2045	Passive	911		
Osgood Sem. Diff: Potency	Strong	1902	Weakness	755		
Osgood Sem. Diff: Evaluation	Positive Outlook	1045	Negative Outlook	1066		
Pleasure, Pain, Vice, Virtue	Pleasure, Feeling, Virtue	936	Pain, Arousal, Vice	1105	Emotion Words	311
Emotional Expressiveness	Overstatement	696	Understatement	319		
Words of Motivation	Means, Persistence	389	Failure	137	Need, Try, Goal	199
Negations & Interjections	'Yes'	20	No/Negation	224	Interjections	24

The other 20 or so categories include aspects of what could be regarded as world knowledge (circa 6000 words) – specialist terms from many disciplines and enterprises, social relations, categories and roles, attributes of human beings, and keywords for referring to objects. Then there is a major category of cognitive orientation comprising over 3500 words and categories that broadly contain tags that referring to grammatical categories of verbs, adjectives and pronouns (Table 2c comprises the details of the 12 other categories):

Table 2c: The distribution of world knowledge, linguistic knowledge, and cognitive orientation tagged words in the *Harvard Dictionary*

Categories	**Subcategories**	
Cognitive Orientation	Knowing, Quality, Quantity	3614
Special (Institutional) Languages	Science, Arts, Humanities, Sports	2800
Verbs	Interpretations, Description	2589
Roles, collectivities, rituals, & interpersonal relations	Social Relations & Roles at work	1732
Communications Processes	Media, Formats	1311
Ascriptive social categories & references to people and animals	Race, Kinship, Gender	1056
Adjective	Attributes of people	754
References to objects	Objects	661
Process and Change	Movement	632
Process and Change	Process	434
References to places, locations and routes between them	Places	312
Pronouns	Personal, Names, Nations?	108

Affect Words, Connotative Meaning and Sentiment Analysis

In my analysis reported in this paper, I have only used the *valence category* words – I have looked at the frequency of *positive* and *negative* words in this category to quantify the attitude that may have been expressed in my NYT corpus towards *Islam* and/or *Muslims.*

Looking at a diachronic analysis, it becomes clear that the frequency of words that have been tagged as words expressing positive evaluation are more frequent than is the case for words that are used to express negative evaluation. Furthermore, the valence words typically comprise less than 10% of all the tokens in the three years that have been studied – namely 1991, 1999 and 2007. In all the three periods, the variation in the frequency of the valence words per month follows the total number of words in the stories where the valence words are found. The number of positive valence words is invariably higher than the negative words (Figures 2a-c).

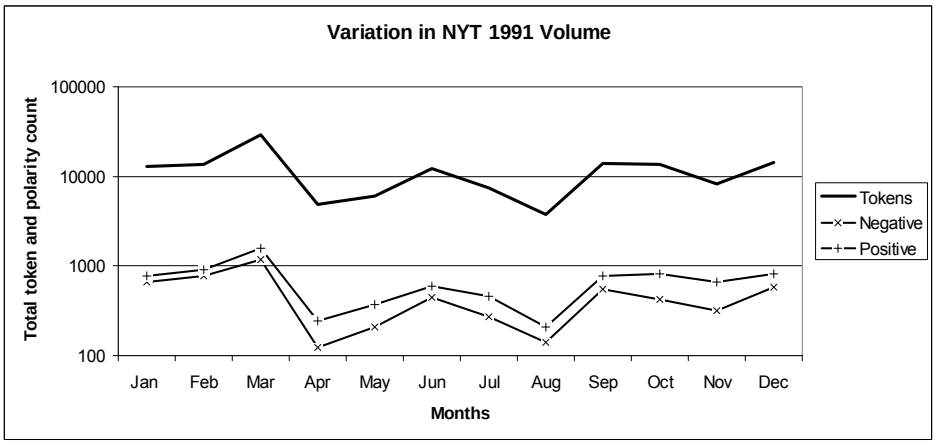

Figure 2a

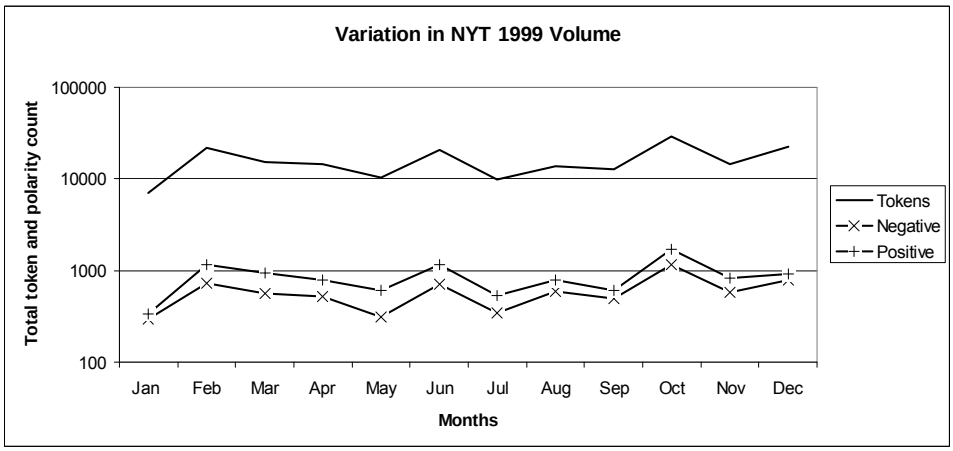

Figure 2b

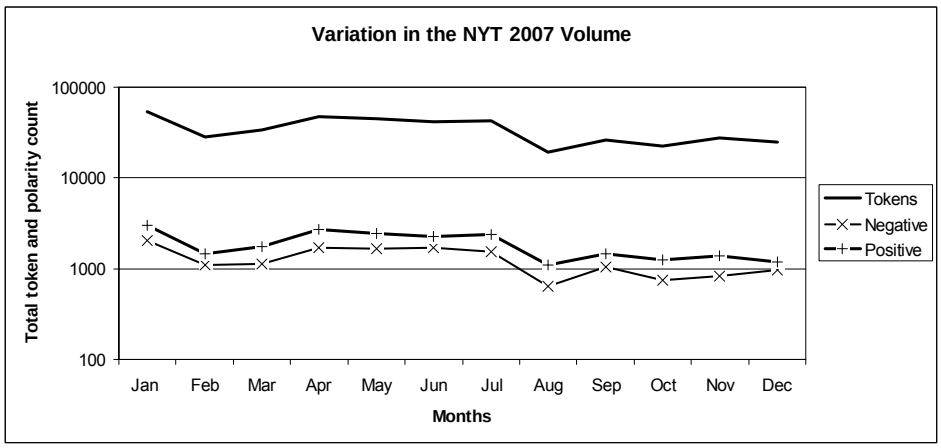

Figure 2c

Figure 2. Variation in the (relative) frequency of valence category terms used to express polar opinions; Figure 2a for the 1991 NYT articles comprising *Islam* AND *Muslim*; Figure 2b *ditto* for 1999; and Figure 2c for 2007

Let me remind the reader that there is an oscillation in the number of positive and negative valence words usually following the total number of words in a given month (Figure 2). The computation of the return values, that is the (logarithm of) the change in the values of the positive and negative affect words from one month to the next, shows that the changes, or returns, in the negative polarity words are, on the whole, more nuanced than is the case for

the returns of positive polarity words. This statement is true for two of the three periods for which I had looked at the NYT stories −1991 and 2007 (see Figures 3a and 3c). The 'picture' for 1999 is confused in that the return values for both positive and negative sentiments have a similar degree of fluctuation (Figure 3-b).

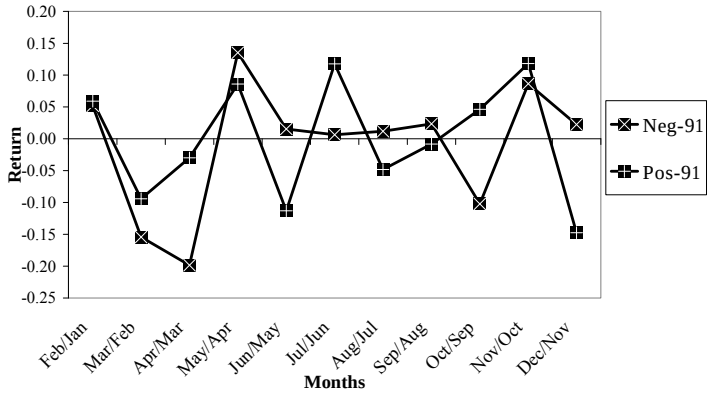

Figure 3a

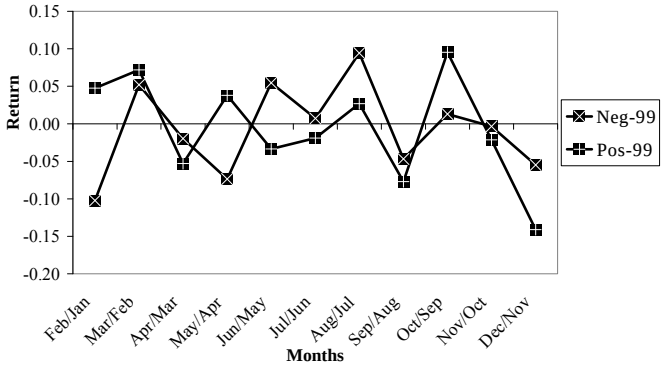

Figure 3b

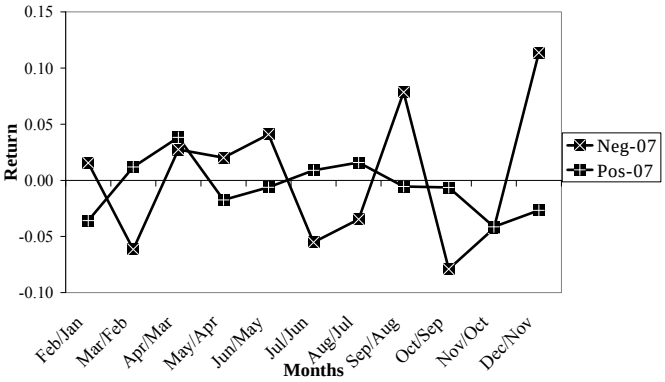

Figure 3c

Figure 3. Variation in the *return values* of valence category terms used to express polar opinions. Figure 3a for the 1991 NYT articles comprising *Islam* AND *Muslim*; Figure 3b *ditto* for 1999; and Figure 3c for 2007

Return and Volatility and Sentiment Analysis

The variation in return values of prices, or volatility of prices of financial instruments, is one of the main concerns in economics and in financial studies (see Taylor 2005 for example). Recall that volatility is defined as the standard deviation of the return values over a fixed period of time – so greater the volatility then greater is the fluctuation in the values of the returns. There are a number of ways in which volatility is defined in the economics and finance – however, there is some consensus that there are *volatility clusters* in the price data, that is, periods of high volatility are followed by high volatility and those of low volatility follow low volatility. The periods when there are rapid changes in prices (periods of *high volatility*) of financial instruments (shares, bonds, or derivatives) are periods where the stakeholders in a market are attempting to discover the 'true' price of a financial instrument. Contrariwise, a period of low volatility indicates some degree of price stability during the period. High volatility indicates changes in attitudes of individuals towards a financial instrument, because of the uncertainty that surrounds the value of the given financial instrument.

Our data is very sparse, monthly rather than minute-by-minute or hour-by-hour data used in economics and finance research, yet both for positive and negative valence word returns we do see some areas of stable returns: there are periods of consecutive low returns for negative polarity words during June-October 1991 surrounded by rapid changes between January to June and November-December 1991 (Figure 3 a). Return values for positive polarity words in 1991 do fluctuate and show a cyclical pattern in 1991 – hence the volatility is high for positive polarity words in 1991. The changes in return values in 1999 (and in 2007) show the 'flat-lining' of the changes in positive polarity words, but the negative polarity words are fluctuating substantially, especially in 2007 (Figure 3c).

I have computed volatility for each of the three years, and for both polarities, over an interval of six months (January to June, and July-December). Usually, volatility of negative values is greater than or equal to positive polarity words, although there are exceptions to this asymmetry in the second half of 1991 and 1999 for positive polarity words. See Table 3 for details.

Table 3. Volatility of positive and of negative polarity words in the NYT sub-corpora (1991, 1999 and 2007)

Period	**Volatility**					
	Neg-91	Pos-91	Neg-99	Pos-99	Neg-07	Pos-07
Jan-Jun	0.14	0.10	0.07	0.05	0.04	0.03
Jul-Dec	0.06	0.10	0.05	0.08	0.08	0.02

Afterword

The use of two well-established measures of unanticipated changes in financial markets – *return* and *volatility* – was presented in this paper for measuring changes in polarity of sentiments expressed about a specially targeted community. Diachronic analysis of texts throws some light on the sentiments that may have been expressed about a group of people or about a system of beliefs. The return and volatility of the negative affect words is worthy of note here.

The views expressed in this paper and the method that has been followed, especially for sentiment analysis, require considerable refinement both in terms of the data used and then the algorithm used in the analysis. The thesaurus of affect words used in this paper, *Harvard Dictionary*, was created in the 1990's and has a much longer legacy; the context of the use of the keywords and of the affect words needs careful analysis; and, remember that the tags used in the description of affect words are ad hoc in some respects.

The traumatic events, that started with the *End of History* onto the *War on Terror* and thence to even more unexpected rise of the *rapidly developing economies*, perhaps have led to the use of sentiment analysis for studying political and religious beliefs of individuals as expressed in a range of typologically distinct languages – Chinese, Arabic, Hindi, Urdu, Russian for instance. And, the text types being analysed comprise not only news and current affairs output about and from targeted communities, but also text and speech fragments found in surreptitiously obtained e-mails and in eavesdropped phone conversation respectively. The results obtained from conventional sentiment analysis should be viewed with some caution, especially where dictionaries of affect may not exist for a given language and where sentiments may be articulated through strategies that are culture-specific.

Acknowledgement

I would like to thank Ann Devitt, formerly my research assistant and now a lecturer at Trinity College, Dublin, Ireland, who wrote a program to compute the distribution of affect words in a text corpus. This program was written under the sponsorship of a collaborative research project on sentiment analysis at Trinity. My thanks to my collaborator, Colm Kearney, Professor of International Business at Trinity College's School of Business Studies, for his contributions and support. Carl Vogel and Donal Holland read the paper in its final stages and my thanks to them for pointing out errors and omissions. I am very grateful to my editor, Ingrid Simonnæs, for helping me to revise the paper and for inserting missing verbs and for discovering phantom tables and figures. The remaining errors, omissions, phantoms and missing verbs that may still lurk around are despite the best efforts of my colleagues. I remain responsible for the contents, informative and distractive, and apologies in advance if I have misled the reader.

References

Ahmad, Khurshid (2007a) Artificial Ontologies and Real Thoughts: Populating the Semantic Web? In Basili, Roberto /Pazienza, Maria Teresa (eds.) (Invited Talk at the) *Annual Conference of Italian Association of Artificial Intelligence. AI*IA 2007, LNAI 4733*. Berlin/Heidelberg: Springer-Verlag. 3–23.

Ahmad, Khurshid (2007b) Being in Text and Text in Being: Notes on representative texts. In Anderman, Gunilla /Rogers, Margaret (eds.) *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters. 60-94.

Ahmad, Khurshid (2008) The 'return' and 'volatility' of sentiments: An attempt to quantify the behaviour of the markets? In Ahmad, Khurshid (ed.) *Proceedings of EMOT 2008: Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology. Workshop at the 13th Language Resources and Evaluation Conference, 27 May 2008*. Marrakesh, Morocco.

Ahmad, Khurshid /Rogers, Margaret (1992) Translation and Information Technology: The Translator's Workbench, *ReCALL*, Vol. 6. 3-9.

Ahmad, Khurshid /Rogers, Margaret (2001) Corpus Linguistics and Terminology Extraction. In Wright, Sue-Ellen /Budin, Gerhard (eds.). *Handbook of Terminology Management. Vol. II. Application-oriented Terminology Management*. Amsterdam /Philadelphia: John Benjamins. 725-760.

Brekke, Magnar /Myking, Johan /Ahmad, Khurshid (1996) Terminology Management and Lesser-Used Living Languages: A Critique of the Corpus-Based Approach. *TKE'96: Terminology and Knowledge Engineering. Proceedings of 4th International Congress on Terminology and Knowledge Engineering, Vienna. (26-28 Aug. 1996)*. Frankfurt: INDEKS-Verlag. 179-189.

Engle, Robert, F. (2003) *Econometric models and financial practice*. http://nobelprize.org/nobel_prizes/economics/laureates/2003/engle-lecture.html (Website accessed 11th August 2008).

Kennedy, Alistair /Inkpen, Diana (2006) Sentiment classification using contextual valence shifters. *Computational Intelligence*. 22 (2). 117-125.

Kugler, Marianne /Ahmad, Khurshid /Thurmair, Gregor (1995) (eds.) *Translator's Workbench: Tools and Terminology for Translation and Text Processing.* Berlin: Springer.

Lasswell, Harold D. (1948) *Power and personality*. London: Chapman & Hall.

Namenwirth, Zvi /Lasswell, Harald D. (1970) *The changing language of American values: A computer study of selected party platforms*. Beverly Hills (Calif.): Sage Publications.

Smadja, Frank (1994) Retrieving collocations from text: Xtract. In Armstrong, Susan (ed.). *Using Large Corpora*. Cambridge, Mass.: MIT Press. 143–177.

Stone, Philip J./ Dunphy, Dexter C. /Smith, Marshall S. /Ogilvie, Daniel M. (1966) *The General Inquirer: A Computer Approach to Content Analysis.* Boston: The MIT Press. (For downloading the GI Lexicon, please go to http://www.wjh.harvard.edu/~inquirer/, Website accessed 16th July 2008).

Taylor, Stephen J. (2005) *Asset Price Dynamics, Volatility, and Prediction*. Princeton and Oxford: Princeton University Press.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62 (3). 1139-1168.