Sentiment in German-language News and Blogs, and the DAX

Robert Remus[1,2], Khurshid Ahmad[1], and Gerhard Heyer[2]

[1] School of Computer Science and Statistics,
Trinity College Dublin, Ireland,
{remusr, kahmad}@cs.tcd.ie
[2] Fakultät für Mathematik und Informatik,
Universität Leipzig, Germany,
heyer@informatik.uni-leipzig.de

Abstract. An analysis of a diachronically organised corpus of German-language newspaper articles and blog posts on economy and finance is presented using a prototype dictionary of affect in German. The changes in the frequency of occurrence of positive and negative polarity words are rendered as return time series and the properties of this time series are described. The *returns* and the variance of returns show that the time series of affect words share a number of properties with that of financial time series in general and in particular with a contemporaneous time series of the DAX – the German stock exchange index. Both the series have outliers not predicted by a normal distribution. Our analysis is a starting point for studying the impact of news and blogs on financial instruments and vice versa.

1 Introduction

1.1 Preamble

It was common wisdom until very recently that decisions taken and prices paid in acquiring financial instruments – shares, bonds, derivatives – are made by rational beings. The use of the term "rational" was meant to exclude emotions, feelings and sentiments. Economists and finance experts claimed even if some of the beings behaved irrationally, the impact of this sentiment-laden behaviour will be countermanded by the large majority of rational agents. The views expressed in news, especially reports of feelings and sentiment of various stakeholders in the financial markets, appear to have had an impact on the markets (Table 1).

The failure of the banking system in 2007/8 indicated that sentiment was or is playing a major role in this fiasco. Speculations, some genuine and others unfounded, about the "health" of financial institutions led to a catastrophic fall in the prices of shares of these institutions; the failure was catastrophic because the so-called efficient market theories, the basis of the orthodoxy that market stakeholders are essentially rational, could not predict the occasional, but persistent wild fluctuation in prices. This is not to say that nobody predicted these wild fluctuations: Benoit Mandelbrot, well-known for his original

Table 1. Reports of sentiment in a German newspaper (*Süddeutsche Zeitung*)

Date	News headline	Exemplar sentence
21.11.2008	Börsenkurse unter Druck: Schwere Übung	"Die *Emotion* steuert das Verhalten, gerade beim Thema *Geld*."
22.10.2008	Finanzmärkte: Aus dem Infekt wird ein Infarkt	"Die *Kapitalmärkte* ticken nicht mehr mit Verstand, sondern nur noch mit *Emotion*."
04.10.2008	Lehren aus der Finanzkrise: Die Mär vom Markt	"Der *Homo oeconomicus* entpuppt sich plötzlich als ein Wesen, das nicht rational handelt, sondern *Gefühle* zeigt."

contribution to fractal geometry of nature, did suggest some 50 or so years ago that the price changes will exhibit this wild behaviour and was able to show it by demonstrating that changes are not distributed according to the widely used *normal distribution*. We use Mandelbrot's approach to investigate the relationship between changes in affect content in news articles, and blogs about financial instruments and institutions in Germany, and changes in the Deutscher Aktien IndeX (DAX).

There is an increasing number of economists and finance experts who hold the view that behaviour of individuals does play a key role in determining prices and preferences. (Engle 2004, Engle and Ng 1993) noted that the news announcements have an impact and that negative "news" has a longer lasting impact than the positive. (Ramchander et al. 2008) have looked at the effect of macroeconomic news on German bonds. The study of "news" in the literature on finance and econometrics relates to the so-called news proxies – the values of key economic and financial indicators and/or the timing of their announcement. It is only recently that an analysis of newspaper articles was undertaken concurrently with the study of the changes in the value of financial assets (Tetlock et al. 2008). We present a method for the analysis and interpretation of German texts of two kinds – newspaper articles and blog posts about Wirtschaft/Ökonomie (economy and finance).

Much of the work in sentiment analysis is Anglo-centric and relies on the existence of affect thesauri which were compiled a few decades ago (Devitt and Ahmad, 2008). There are promising developments in the area of modern European languages, especially German and French, where digital dictionaries of metaphors are being developed; the long-running and ambitious *EuroWordnet* (EWN) Project is starting to create metaphor databases together with national initiatives in Germany, particularly the work carried out under the rubric of Hamburg Database project is relevant here (Eilts and Lönneker 2002, Lönneker and Eilts 2004). The use of WordNet's template imposes its own limitations as far as metaphorical language is concerned. Over the last five or so years the Hamburg team has collated and annotated over 1500 examples of the use of non-literal language in form of a corpus of texts in German and French (Lönneker-Rodman 2008).

A detailed and validated dictionary of metaphorical words is critical for sentiment analysis, however such resources are in planning or development stages. We have resorted to an *intuitive* approach and constructed our own glossary of affect words in German.

1.2 A German Case Study

We had collected a corpus of newspaper articles and blogs on economy and finance over a three year period and compared the changes in affect content in the corpora with the changes in the DAX. Such a study may throw some light on whether the changes in affect content and prices have any correlation. Such a correlation may, in turn, enlighten us about a possible impact of prices on news and blogs and/or vice versa. First, the details of the text corpora is given, followed by a description of the *ad hoc* German affect dictionary created intuitively, and finally the description of the DAX time series is presented.

News and Blogs. We explore possible answers to our research questions by analysing candidate sentiments as expressed in economic and financial news sections of a quality German newspaper, the *Süddeutsche Zeitung* (a corpus of 3.91 million words) and potential sentiments as articulated in four blogs focused on financial trading in Germany (a corpus of 0.43 million words). The news articles and blog posts were published and posted respectively in 2006–2008.

German Dictionary of Affect. Our study uses Harvard University's General Inquirer (GI) lexicon as described in (Stone et al. 1966), and the words categorized as "Pos" (1,914 terms) and "Neg" (2,293 terms), which have positive and negative semantic orientations respectively in particular. These terms and the terms identified using a *local grammar* by (Ahmad, Cheng and Almas 2006) were translated into German by a mixture of human and machine translation. In this way a total of 9,301 positive and 10,697 negative German polarity candidate word forms were identified. These word forms contain inflections of a given lemma as well.

The frequency of occurrence of negative and positive polarity candidates follows a Zipf-like distribution (Zipf 1949): few candidates are repeatedly used, some less so and a large number are extremely infrequent – the *hapex logomena*. The distribution of the frequent words remains constant and the overall contribution of these candidates to the text corpora (viewed annually) remains constant at around 4% for positive words (see Table 2) and between 2–3% for negative words (see Table 3).

The most frequent positive candidates are *viel* and *viele* – these are the equivalent of the English *plenty*, which has been entered as a "positive" affect word in the GI lexicon; similarly *gegen* our most frequent negative polarity candidate is the equivalent of the English *against*, which has been entered as a "negative" affect word in the GI lexicon. Table 3 shows how the frequency correlates with aspects of German and global economy – while *krise* (crisis) was amongst the

Table 2. The annual distribution of the most frequent positive affect words in our newspaper corpus. N is corpus size, f is absolute frequency, $N_{06} = 673,592$, $N_{07} = 1,318,333$, $N_{08} = 1,919,179$

		2006			2007			2008	
Rank	Token	f	$\frac{f}{N_{06}}$	Token	f	$\frac{f}{N_{07}}$	Token	f	$\frac{f}{N_{08}}$
1	viele	511	*0.08%*	viele	1074	*0.08%*	viele	1727	*0.09%*
2	viel	491	*0.07%*	viel	931	*0.07%*	viel	1518	*0.08%*
3	gut	436	*0.07%*	gut	898	*0.07%*	gut	1382	*0.07%*
4	grossen	348	*0.05%*	angebot	702	*0.05%*	macht	1056	*0.06%*
5	macht	324	*0.05%*	grossen	688	*0.05%*	grossen	1016	*0.05%*
6	grosse	307	*0.05%*	macht	651	*0.05%*	grosse	894	*0.05%*
7	teil	261	*0.04%*	geben	623	*0.05%*	geben	781	*0.04%*
8	geben	237	*0.04%*	grosse	573	*0.04%*	teil	778	*0.04%*
9	angebot	215	*0.03%*	teil	450	*0.03%*	erhalten	639	*0.03%*
10	erhalten	207	*0.03%*	erhalten	393	*0.03%*	angebot	585	*0.03%*
	Top 10	3,337	*0.50%*		6,983	*0.53%*		10,376	*0.54%*
	Total	29,289	*4.35%*		56,367	*4.28%*		83,587	*4.36%*

top 10 words in 2006, this has been replaced by the more informative *finanzkrise* (financial crisis) in 2008. Also it appears that the occurrence and/or threat of *streik/streiks* (strikes) has decreased between 2007 and 2008. The "mood" of the news reports is distinctly turning *negative* if the basis was frequency count of negative words – which rose by around 40% in three years versus a small 4% increase in positive words.

The distribution of positive and negative words in our blog corpus is not exactly the same as that of the more formal newspaper corpus (see Table 4), but there is a sharing of key word forms: *viel(e)*, *gut* and *ende*, *krise*, and *gegen*.

DAX. We use the Deutscher Aktien IndeX 30 (DAX 30), the German Stock Index, that comprises the 30 largest and most actively traded german companies admitted to the Prime Standard segment, which are listed in the Frankfurt Stock Exchange. As the DAX 30 changes over the day, the closing values were used for a period of three years, 2006–2008.

The near-global sub-prime mortgage crisis of 2007 and the subsequent bank "bail-outs" or "quantitative easing" in 2008, has had its toll on the DAX and the variation in its rate of change has become more pronounced as we move from 2006 – small positive and negative changes – onto almost wild swings in 2008. The key concept of *return* (r_t) plays a major role here; this variable is defined as the natural logarithm of the ratio of the value of the DAX today and yesterday:

$$r_t = \ln\left(\frac{\mathrm{DAX}_t}{\mathrm{DAX}_{t-1}}\right) \tag{1}$$

The wild swings in the returns indicate a reversal of trend and overall behaviour of the time series (see Fig. 1). The returns almost appear as "noise"

Table 3. The annual distribution of the most frequent negative affect words in our newspaper corpus. N is corpus size, f is absolute frequency, $N_{06} = 673,592$, $N_{07} = 1,318,333$, $N_{08} = 1,919,179$

		2006			2007			2008	
Rank	Token	f	$\frac{f}{N_{06}}$	Token	f	$\frac{f}{N_{07}}$	Token	f	$\frac{f}{N_{08}}$
1	gegen	907	*0.14%*	gegen	1975	*0.15%*	gegen	3140	*0.16%*
2	ende	521	*0.08%*	ende	1132	*0.09%*	ende	1687	*0.09%*
3	fall	286	*0.04%*	streik	976	*0.07%*	finanzkrise	1377	*0.07%*
4	kosten	246	*0.04%*	streiks	817	*0.06%*	krise	1254	*0.07%*
5	knapp	234	*0.04%*	fall	578	*0.04%*	fall	1002	*0.05%*
6	trotz	206	*0.03%*	kosten	492	*0.04%*	knapp	704	*0.04%*
7	kritik	157	*0.02%*	knapp	416	*0.03%*	kosten	694	*0.04%*
8	konkurrenz	151	*0.02%*	streit	378	*0.03%*	trotz	642	*0.03%*
9	problem	146	*0.02%*	trotz	342	*0.03%*	streik	627	*0.03%*
10	krise	137	*0.02%*	affaere	339	*0.03%*	streit	608	*0.03%*
	Top 10	2,991	*0.44%*		7,445	*0.57%*		11,735	*0.61%*
	Total	17,646	*2.62%*		36,389	*2.76%*		83,587	*3.14%*

accompanying the main signal (the close values) until January 2008 despite the 45% increase in the value of the capitalization of German firms that make up the DAX. However, returns look more spikey between January and February 2008 and the last quarter of 2008 shows very large spikes; larger spikes are notated as *variance breaks* – a point in which markets show an about turn from its previous behaviour – in the econometric literature (Taylor 2005) and in time series literature (Gençay, Selçuk and Whitcher 2002).

The variance break during the end of 2008 is demonstrated clearly by a substantial change in the *volatility* (standard deviation of returns) of the DAX: the quarterly volatility is between 0.01 and 0.015 and jumps to 0.02 (a factor of two increase in three months) and by the end of 2008 the volatility jumps to 0.04 (see Fig. 1). We conducted a similar exercise on the sum of positive and negative affect words, the total affect, occurring in our newspaper corpus and our blog corpus (see Figs. 2 and 3). The changes in the returns and volatility are much more pronounced in the total affect returns for news articles and blog posts when compared with the DAX returns and volatility. The key difference here is that volatility in both news articles and blog posts decreases on the onset of the crisis in the DAX at the end of 2008.

2 Method: the Return Series

It has been argued that the impact of news and blogs on financial markets may be computed as the variation in affect word distribution in news and blogs with that of key financial market indices (Ahmad 2008). The motivation in making such a comparison is this: the behaviour of the returns in a large number of share prices and market indices, currency exchange rates, bonds, commodities

Fig. 1. Daily return and quarterly volatility in the DAX (2006–2008)

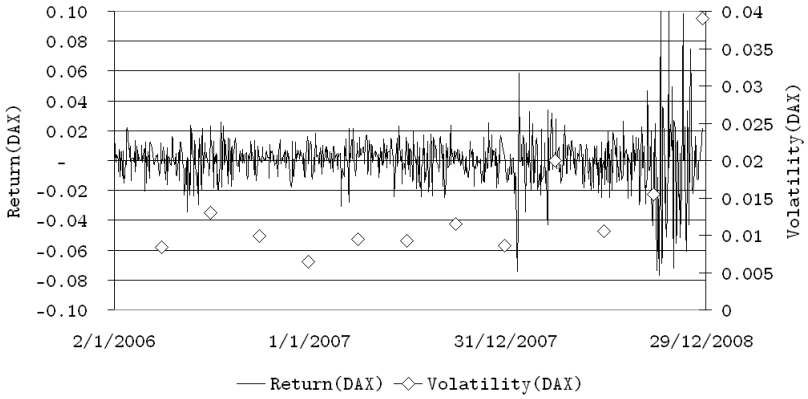

Fig. 2. Daily return and quarterly volatility in the total affect of the news (2006–2008)

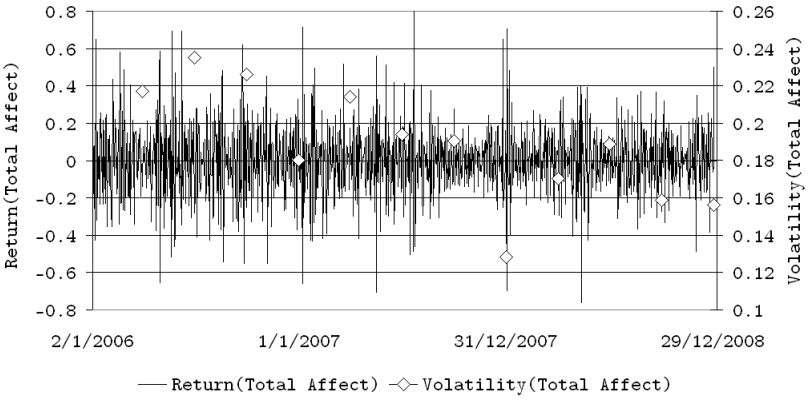

Fig. 3. Daily return and quarterly volatility in the total affect of the blogs (2006–2008)

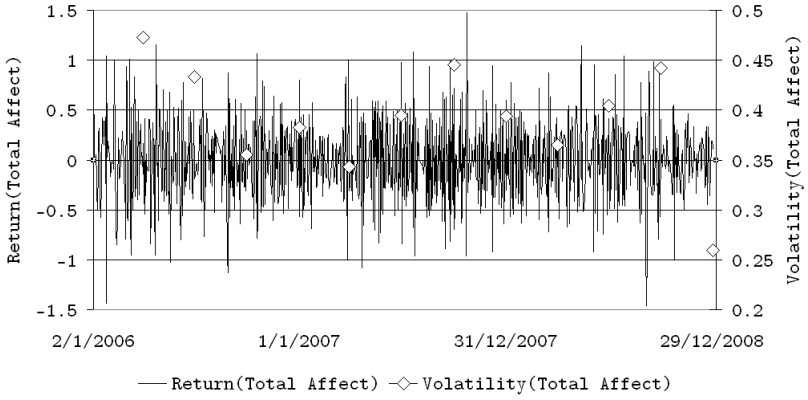

Table 4. The distribution of affect words in our blog corpus (2006 – 2008). N is corpus size, f is absolute frequency, $N = 451,520$

	Positive affect words			Negative affect words		
Rank	Token	f	$\frac{f}{N}$	Token	f	$\frac{f}{N}$
1	gut	820	*0.18%*	ende	454	*0.10%*
2	kaufen	676	*0.15%*	leider	294	*0.07%*
3	steigen	578	*0.13%*	unten	224	*0.05%*
4	gewinne	509	*0.11%*	fallen	219	*0.05%*
5	korrektur	490	*0.11%*	crash	194	*0.04%*
6	gewinn	450	*0.10%*	trotz	134	*0.03%*
6	viel	450	*0.10%*	fall	128	*0.03%*
7	viele	418	*0.09%*	krise	128	*0.03%*
8	wert	341	*0.08%*	langsam	126	*0.03%*
9	aktuell	338	*0.07%*	gegen	125	*0.03%*
10	gute	335	*0.07%*	kurz	124	*0.03%*
	Top 10	5,405	*1.19%*		2,150	*0,48%*
	Total	25,477	*5.64%*		9,361	*2.07%*

(ranging from precious metals to livestock), appears quite similar at a given level of statistical description. This description relates to the "general properties that are expected to be present in any set of returns" (Taylor 2005:51). If the time variation of affect words, both positive and negative have the same general properties as that of, say, financial time series, then there will be a basis of comparing, contrasting and correlating the distribution of affect words over an interval of time with that of stylised variables of a financial time series.

The behaviour of the returns is studied in finance and econometrics: it has been argued that the probability of rise of prices, or positive return, and fall, or negative return, will be almost equal. The returns, the argument runs, will be distributed in a "bell curve", expressed for example through a normal distribution, with greater probability of smaller returns and lesser probability of higher returns – with unusually high returns being statistically improbable. The symmetric "bell curve", it turns out, does not quite exist because there is a history of boom and bust in financial markets, and one can extrapolate that this happens due to the occurrence of these improbable returns (Mandelbrot and Hudson 2004).

(Taylor 2005) has suggested that one should use different moments of the distribution of a time series to compile "summary statistics" – the first moment being the mean of the series, the second relates to the standard deviation, the third to the skewness and the fourth to the kurtosis. The standard deviation is also used to compute the *volatility* of a time series over different intervals within a set of observations. (Taylor 2005) goes on to suggest that

1. One key property of the time series of the various assets is that the distribution of their returns is not normal (Taylor 2005:69),

2. the returns are characterized by three key properties: their distributions are approximately symmetric, have fat tails, and have a high peak (Taylor 2005:69),
3. the returns show no correlation between returns for different days (Taylor 2005:77) and
4. there is a positive dependence between absolute returns on nearby days, and likewise for squared returns (Taylor 2005:82).

In this paper we will focus on properties 1 and 2 for the time series of affect words found in news articles and blog posts. Lack of space and comparatively short run of our time series (3 years of daily data rather than 10 years as the case in Taylor) does not allow us to discuss properties 3 and 4.

There are ontological questions related to our affect "time series": first, is affect a time-ordered phenomenon that can be organized on a discrete grid? And second, maybe the more pertinent question, how accurate is a frequency count of words, independent of context, in measuring something as elusive as affect? But for us our frequentist, i.e. word counting approach is the beginning of a more complex study – an exploration if you will.

3 Results

3.1 Stylised Variables

The stylised variables have been tabulated according to (Taylor 2005): The variation in the returns is large for all five series we have considered, being slightly more asymmetric for all the series as evidenced by the minimum and maximum values of the returns. The mean (scaled by 10^4) of the DAX is negative (reflecting the drop in its value towards the end of 2008); the mean of the negative affect series is higher for both news reports and blogs when compared to similar means for positive affect. This difference in the negative and positive affect series is demonstrated in the second moment (standard deviation) as well. The kurtosis values show that the distribution of the DAX is very sharp – reflecting the steady state values of the DAX in the first 30 months of our observation, all returns are clustered around zero. The kurtosis for negative affect is smaller than that for the positive affect (see Table 5).

Table 5. Stylised variables for the five time series reported

Time series	Min	Max	$10^4\times$ Mean	$10^2\times$ Std Dev	Skewness	Kurtosis
DAX	-0.09	0.11	-2.03	1.61	0.15	11.14
News Positive	-1.82	1.66	1.65	28.21	-0.07	5.45
News Negative	-1.54	1.69	8.29	35.44	-0.03	1.56
Blogs Positive	-2.2	2.08	5.34	45.02	-0.02	1.81
Blogs Negative	-2.18	2.48	16.18	68.96	0.04	0.24

3.2 Non-Normal Distribution

The differences between a normal distribution and the actual distribution of returns increases as we move from returns for blogs to that for news and onwards to the DAX. All our series contain outliers beyond the third standard deviation which is very unlikely in a normal distribution. This fact is known for the DAX (Taylor 2005) and we have confirmed this for the news and blogs affect time series (see Table 6).

Table 6. A comparison of the normal probability density function with the distribution of returns for affect content in news, blogs and the DAX

Probabiliy distribution between	Normal	Blogs		News		DAX
		Positive	Negative	Positive	Negative	
0 to 0.25	*19.74%*	19.53%	20.12%	24.74%	22.77%	33.46%
0.25 to 0.5	*18.55%*	19.88%	19.41%	23.61%	20.41%	22.31%
0.5 to 1	*29.98%*	33.14%	31.12%	30.01%	29.82%	27.17%
1 to 1.5	*18.37%*	15.74%	15.74%	11.67%	14.77%	10.24%
1.5 to 2	*8.81%*	7.34%	8.76%	4.70%	7.15%	3.02%
2 to 3	*4.28%*	3.55%	4.38%	3.57%	4.14%	1.44%
3+	*0.27%*	0.83%	0.47%	1.69%	0.94%	2.36%
	100%	100%	100%	100%	100%	100%

4 Conclusion

This paper presents the initial results of a joint project between Trinity College, Ireland and University of Leipzig, Germany. Our study shows that a return series of affect words, extracted either from newspapers or blogs, has some characteristics of a financial time series. Furthermore, our approach based on the use of returns volatility helps us to make comparisons between time series that have different units of measurements.

We are improving the coverage of our sentiment dictionary and currently a human evaluation of the affect words, shown as keywords in context, is underway. We are also increasing the time span of our data and the intention is to re-run all the calculations on a 10-year data set, rather than a 3-year data set. Once our dictionary is validated and the coverage in terms of time expanded, we will be able to confirm the findings of this paper and draw further conclusions.

References

Ahmad, Khurshid: The "Return" and "Volatility" of Sentiments: An Attempt to Quantify the Behaviour of the Markets? In Proc. of the Workshop Emotion, Metaphor, Ontology & Terminology of the 6th Intl. Conf on Language Resources and Evaluation (2008)

Ahmad, Khurshid, David Cheng, and Yousif Almas: Multi-lingual Sentiment Analysis in Financial News Streams. In Proc. of the 1st Intl. Conf. on Grid in Finance (2006)

Devitt, Ann, and Khurshid Ahmad: Sentiment Analysis and the Use of Extrinsic Datasets in Evaluation. In Proc. of the 6th Intl. Conf on Language Resources and Evaluation (2008)

Eilts, Carina, and Birte Lönneker: The Hamburg Metaphor Database. Metaphorik.de, 3/2002 (2002)

Engle, Robert F.: Nobel Lecture. Risk and Volatility: Econometric Models and Financial Practice. American Economic Review. Vol. 94, Pages 405–420 (2004)

Engle, Robert F., and Victor K. Ng: Measuring and Testing the Impact of News on Volatility. Journal of Finance. Vol. 48, Pages 1749–1777 (1993)

Gençay, Ramazan, Faruk Selçuk, and Brendan Whitcher. An Introduction to Wavelets and Other Filtering Methods in Finance and Economics. London: Academic Press (2002)

Lönneker, Birte, and Carina Eilts: A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information. In Proc. of the 2nd Global Word Net Conference, Pages 157–162 (2004)

Lönneker-Rodman, Birte: The Hamburg Metaphor Database project: issues in resource creation. In Proc. of Language Resources and Evaluation, Vol. 42, Pages 293–318 (2008)

Mandelbrot, Benoit, and Richard L. Hudson: The (Mis)Behaviour of Markets – A Fractal View of Risk, Ruin and Reward. New York: Basic Books (2004)

Ramchander, Sanjay, Marc W. Simpson, and Harold Thiewes: The Effect of Macroeconomic News on German Closed-end Funds. The Quarterly Review of Economics and Finance. Vol. 48, Pages 708–724 (2008)

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie et al.: The General Inquirer: A Computer Approach to Content Analysis. Boston: The MIT Press (1966)

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

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy: More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance. Vol. 63, Pages 1437–1467 (2008)

Zipf, George K.: Human Behaviour and the Principle of Least Effort. Cambridge: Addison-Wesley Press (1949)