

**Emotional News: How Emotional Content of News and
Financial Markets are Related**

by

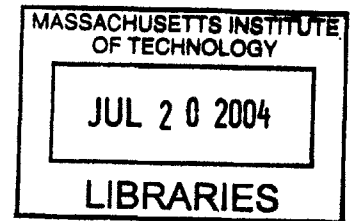
Wan Li Zhu

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degrees of
Bachelor of Science in Electrical [Computer] Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

May 7, 2004

[June 2004]

Copyright 2004 Wan Li Zhu. All rights reserved.



The author hereby grants to M.I.T. permission to reproduce and
distribute publicly paper and electronic copies of this thesis
and to grant others the right to do so.

Author _____
Department of Electrical Engineering and Computer Science
May 7, 2004

Certified by _____
Andrew W. Lo
Harris & Harris Group Professor
Thesis Supervisor

Certified by _____
Dmitry V. Repin
Postdoctoral Associate
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Emotional News: How Emotional Content of News and Financial Markets are Related

by
Wan Li Zhu

Submitted to the
Department of Electrical Engineering and Computer Science

May 7, 2004

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical [Computer] Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

We present here a first step towards developing a quantitative model that relates investor emotions to financial markets. We used Wall Street Journal articles as a proxy of investor emotions on a “macro” level. We measured the emotional characteristic of the article texts quantitatively through content analysis to arrive at a daily set of emotional and subject category scores. After establishing the statistical and informational validity of these scores, we ran correlations and regressions between the daily category scores and broad market indices variables such as return, volume, and volatility to determine whether there is a relationship. We found that negative emotions are more strongly correlated with market variables than positive emotions. We also found that markets are a better predictor of emotions than emotions of markets. There also appears to be a stronger relationship between emotions and market volatility than with market returns. In investigating the source of the correlations, we found that the most extreme category scores are responsible for driving the bulk of the correlations. Event study results suggest that there is a stronger relationship between negative events and negative emotions than between positive events and positive emotions. A challenge we encountered that remains to be fully addressed is how to integrate our interpretation of the analysis results into our understanding of the link between emotions and financial markets from a causal and psychological perspective.

Thesis Supervisor: Andrew W. Lo
Title: Director, MIT Laboratory for Financial Engineering

Acknowledgements

I would like to thank the following people for helping me complete this thesis:

My family

Bei Li, Wei Guang, Diana, Huei Ying
For their love and support

My thesis supervisor

Andrew Lo

Whose vision and inspiration is the reason why
I decided to pursue my graduate studies
Whose patience and guidance was crucial to this research
Many of these ideas are either direct implementations of Andrew's ideas
or had their roots there

My supervisor and mentor

Dmitry Repin

Whose continued advice and hands-on help has allowed me to
overcome many conceptual and technical challenges
Many of these ideas are either direct implementations of Dmitry's ideas
or had their roots there

My colleagues

Mike Epstein, Mila Getmansky, Jasmina Hasanhodzic, Xian Ke
For always being available to discuss problems

Everyone at AlphaSimplex Group

For helpful feedback on the approach and results of this research

My many friends and colleagues in and out of MIT for

Discussion

Criticism

Support

Advice

Love

Table of Contents

Cover Page	1
Abstract	3
Acknowledgements	3
Table of Contents	4
1. Introduction	5
1.1 Motivation	5
1.2 Overview of Approach	5
1.3 Related Work	6
2. Data	7
2.1 Wall Street Journal	7
2.2 Subject & Emotional Category Scores	7
2.3 Market Variables	15
2.4 Event Study	17
3. Methodology	21
3.1 Validity of GI-Extracted Content	21
3.2 Definition of Dimensions	22
3.3 Correlations, Regressions	25
3.4 Event Study	25
4. Results	27
4.1 Correlation Results	27
4.2 Regression Models	34
4.3 Event Study Results	45
5. Discussion	49
5.1 Correlations	49
5.2 Regressions	52
5.3 Event Study	53
6. Future Work	55
6.1 Emotional Category Creation	55
6.2 Phrase-Level Textual Processing	55
6.3 Subject-Specific News	56
6.4 Emotional Index	57
7. Conclusion	58
8. References	60
Appendix A: The Wall Street Journal	61
Appendix B: Emotional Category Creation	62

1. Introduction

1.1 Motivation

Financial literature is filled with research that demonstrates the impact of investor emotions on financial markets. With data from twenty-six international stock exchanges, Hirshleifer and Shumway (2003) suggested that good moods resulting from morning sunshine lead to higher stock returns, linking investor optimism with stock performance. Daniel, Hirshleifer, and Subrahmanyam (1998) showed that investor overconfidence leads to negative long-lag autocorrelations and excess volatility in the securities market. Lee, Shleifer, and Thaler (1991) confirmed that closed-end fund discounts form a measure of individual investor sentiments, which act to make funds riskier than the portfolio they hold and causes average underpricing of funds relative to fundamentals.

While research has indicated the contribution of investor sentiments to financial markets, a quantitative measure of general investor emotions and their relationships to financial markets have yet to be discovered and validated. A quantitative model would provide us with a more concrete and actionable understanding of the relationship between investor emotions and financial markets. The closest such quantitative model we have today is the CBOE Market Volatility Index (VIX), now widely referred to as the market's fear indicator. The VIX measures the implied volatility of the U.S. equity market such that when markets decline, the VIX index usually moves inversely, rising to reflect the increase in demand for puts. It is hypothesized that a higher VIX value implies a higher level of fear in the market. However, since the VIX is computed directly from the S&P100 index option prices, it is a market-dependent measure of investor emotions. Therefore, it wouldn't be able to tell us if certain investor emotions are not reflected in the market. To achieve this, we need a measure of investor emotions that is not directly derived from market prices.

1.2 Overview of Approach

The challenge of developing a quantitative model of investor emotions is two-fold. First, there is no one generally accepted set of emotion definitions. Second, investor emotions manifest themselves in many different forms that are difficult to quantify. To address those challenges, we propose a novel approach for measuring investor emotions by performing content analysis on the entire daily article text of the Wall Street Journal from 1991 to 2002. We chose news articles as a proxy of general public sentiment and the Wall Street Journal in particular because of its large readership audience and business and finance focus. By counting the occurrence of the words in news articles that have been shown to indicate certain emotions, we arrive at a daily set of scores for each distinct emotional category. Next, we demonstrate the validity of these emotional scores and show their relationships to the broad financial markets. We also perform event study analysis to examine the impact of major events on these category scores.

1.3 Related Work

There has been a significant amount of work done in investigating the relationship between news and financial markets. One research effort that bears resemblance to our work is Niederhoffer's 1971 study of world events and stock prices. World events were defined as five- to eight- column headlines in the New York Times and then organized into categories of meaning. Niederhoffer found that large stock price changes did follow world events more than randomly selected days but that a particular category into which a world event falls did not add much additional information about future price movements. Other findings include a strong tendency for large price changes on the first and second day following world events to show the same direction of change and market overreaction to bad news as indicated by price rises on days 2-5 following extremely bad world events.

Measuring public information by the number of news releases by Reuter's News Service per unit of time, Berry and Howe (1994) showed that there is a positive, moderate relationship between public information and trading volume. Engle and Ng (1993) defined the news impact curve which measures how new information is incorporated into volatility estimates. Hong, Lim, and Stein (2000) confirmed that firm-specific information, especially negative information, diffuses only gradually across the investing public.

By studying the number of news announcements reported daily by Dow Jones & Company, Mitchell and Mulherin (1994) did not find any strong relations between news and market activity. Pearce and Roley (1985) showed that on announcement days, surprises related to monetary policy significantly affect stock prices, but only limited evidence of an impact from inflation surprises and no evidence of an impact from real activity surprises.

2. Data

2.1 Wall Street Journal

Rather than analyzing headlines or any specific type of news such as macroeconomic or firm-specific announcements, we decided to include the entire daily article texts of the Wall Street Journal from January 2, 1991 to December 31, 2002 to arrive at a “macro” measure of emotions. Weekend journals were not included. Both title and body of article texts were included and treated in the same manner in the analysis. All sections of the Journal were included in the analysis, including the section “What’s News” which highlights summaries of articles on a given day. We realize that despite some repetition in news content, the emotional content of these texts may be different. Advertisements and graphical figures of the Journal were not included. We provide some background information on the Wall Street Journal from 1991-2002 in Appendix A.

2.2 Subject & Emotional Category Scores

In order to analyze the Wall Street Journal text, we used a tool called the General Inquirer (GI). GI is basically a mapping tool that maps each text file to counts on dictionary-supplied categories¹. Each category is a list of words and word senses. The currently distributed version combines the Harvard IV-4, the Lasswell, and five categories based on the social cognition work of Semin and Fiedler, making for 182 categories and 11,767 words and word senses in total¹.

Given a plain-text document, GI outputs a list of category scores indicating what percentage of words in the document was found in each dictionary category. For example, consider a text document of three words: “word₁ word₂ word₃”. Suppose “word₁” is found in Category₁, “word₂” in Category₂, and “word₃” not in any of GI’s categories. The computed scores for this text are Category₁ = 33.3% and Category₂ = 33.3%.

GI performs basic word root analysis, so that words such as “happily,” “happier,” and “happy” are all recognized as the same word. One weakness of GI’s word-level analysis

¹ General Inquirer Website, “How the General Inquirer is Used and a Comparison of General Inquirer with other Text-Analysis Procedures,” [Web page document], Available HTTP:

<http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html>

- Harvard IV-4 categories as described on the General Inquirer Website above
- Lasswell categories as described in *Dynamics of Culture* by J. Zvi Namenwirth and Robert Philip Weber. 1987. Winchester MA: Allen & Unwin
- Semin and Fiedler categories as described in *Journal of Personality and Social Psychology*, 1988, 54, 558-568

is that it has no ability to comprehend phrases. For example, the phrase “not happy” is not recognized as one phrase but as the words “not” and “happy” separately. We believe that this weakness will not invalidate our results because individual keywords still account for the bulk of meanings in texts. GI also tries to disambiguate words, so that when it encounters “address,” for example, it will try to determine whether it is “address” the noun or “address” the verb.

We’ve selected a subset of 38 categories for our analysis listed in Table 1. The categories we chose can be divided into three general types:

- Subject content categories: Econ@, Legal, Milit, Polit@, etc.
- Emotional content categories: Arousal, Feel, Pain, Pleasur, WlbPsysc, etc.
- Other categories: Increas, Decreas, Complet, Fail, etc.

While the focus of this study is on the emotional content category scores and their relationships with financial markets, we’ve included other category types for both completeness and control purposes. Distributions of daily category scores as computed by GI on the Wall Street Journal article text from 1991 to 2002 is shown in Figure 1a where all distributions are on the same scale. Figure 1b shows the same distributions on individual scales. Many of the category scores resemble a normal distribution. Table 2 shows the statistical properties of the category scores over the entire period of 1991 to 2002. Category scores show strong stationarity with one period autocorrelations ranging from 20% to 70%.

We ran correlations between the scores of the different categories for the entire period of 1991 to 2002. The results are shown in Table 3. Some emotional categories such as Negativ and Weak exhibit high correlations with each other because they share many words in common. However, certain categories that exhibit strong correlations but do not share many words in common include WlbPsysc and Econ@ (-40%), Exprsv and Econ@ (-50%), and Active and Passive (41%).

Table 1 Selected General Inquirer Categories²

Category Name	Definition	Example Words	Number of Words	Source Dictionary
Active	Words implying an active orientation	Accomplish, celebrate, change, foster, mislead, oust, widen	2,045	Harvard IV-4
AffTot	Words in the affect domain	Care, faithful, home, jealous, loyal, passion, sorrow, zeal	196	Lasswell
Arousal	Words indicating excitation, aside from pleasures or pains, but including arousal of affiliation and hostility	Antagonize, grateful, insistent, motivate	166	Harvard IV-4
Complet	Words indicating that goals have been achieved, apart from whether the action may continue	Attain, comprehensive, fulfill, recover, sustain	81	Harvard IV-4
Decreas	Words indicating decrease, lessening	Diminish, erode, languish, refine, weaken	82	Harvard IV-4

² General Inquirer Website, “Descriptions of Inquirer Categories and Use of Inquirer Dictionaries,” [Web page document], Available HTTP: <http://www.wjh.harvard.edu/~inquirer/homecat.htm>

Econ@	Words of an economic, commercial, industrial, or business orientation, including roles, collectivities, acts, abstract ideas, and symbols, including references to money. Includes names of common commodities in business.	Anti-trust, bankrupt, bid, capital, dollar, fiscal, investment, oil, salary, price, unemployment, valuation	510	Harvard IV-4
EMOT	Words related to emotion that are used as a disambiguation category	Adore, brood, despair, grief, nervous, pride, terror	311	Harvard IV-4
Exprsv	Words associated with the arts, sports, and self-expression	Art, baseball, biography, concert, critic, fashion, medal, sing, vacation	205	Harvard IV-4
Fail	Words indicating that goals have not been achieved	Abandon, disarm, helpless, lapse, mishap	137	Harvard IV-4
Feel	Words describing particular feelings, including gratitude, apathy, and optimism, not those of pain or pleasure	Aloof, fiery, obstinate, qualm, upbeat	49	Harvard IV-4
Goal	Names of end-states towards which muscular or mental striving is directed	Accomplishment, destination, innovation, victory	53	Harvard IV-4
If	Words denoting feelings of uncertainty, doubt and vagueness	Approximate, barely, confuse, maybe, postpone, reluctant, suspicious, unexpected, wary	132	Lasswell
Increas	Words indicating increase, heightening	Accelerate, broaden, elaborate, expand, prosper	111	Harvard IV-4
Legal	Words relating to legal, judicial, or police matters	Accuse, contract, crime, divorce, guilty, indictment, negligence, prison, repeal, verdict	192	Harvard IV-4
Means	Words denoting objects, acts or methods utilized in attaining goals	Access, budget, crucial, facility, make, method, resource	244	Harvard IV-4
Milit	Words relating to military matters	Bomb, commander, fleet, guard, missile, radar, stronghold, weapon	88	Harvard IV-4
Need	Words related to the expression of need or intent	Crave, envy, hope, intent, relish, urge, want	76	Harvard IV-4
Negativ	Words of negative outlook	Abject, belittle, deception, havoc, perplex, resent, stagnant, turbulent	2,291	Newly Constructed
No	Words directly indicating disagreement, with the word "no" itself disambiguated to separately identify absence or negation	Disagree, nay, no, nope, wrong	7	Harvard IV-4
Ovrst	Words indicating emphasis in realms of speed, frequency, causality, inclusiveness, quantity or quasi-quantity, accuracy, validity, scope, size, clarity, exceptionality, intensity, likelihood, certainty and extremity	Accentuate, alarming, bulk, chaos, dominant, emphasis, hopeless, inevitable, notable, perpetual, severe, unique	696	Harvard IV-4
Pain	Words indicating suffering, lack of confidence, or commitment	Agony, discomfort, dismay, downfall, hysteria, sad, weary	254	Harvard IV-4
Passive	Words indicating a passive orientation	Admit, coincide, depend, hesitant, lack, lost, reflect, trust, worsen	911	Harvard IV-4
Persist	Words indicating "stick to it" and endurance	Always, deadlock, incessant, prolong, unflinching	64	Harvard IV-4

Pleasur	Words indicating the enjoyment of a feeling, including words indicating confidence, interest and commitment	Admire, celebrate, confident, delight, grateful, relief, upbeat	168	Harvard IV-4
Polit@	Words having a clear political character, including political roles, collectivities, acts, ideas, ideologies, and symbols	Alliance, campaign, civil, congress, elect, freedom, legislation, tariff, treaty, vote	263	Harvard IV-4
Positiv	Words of positive outlook	Accept, advance, confident, discreet, favorite, ideal, natural, realistic, solution, upbeat	1,915	Newly Constructed
ReEthic	Words of values concerning the social order	Adhere, fair, faith, goodwill, indignant, moral, offence	151	Lasswell
RspTot	Words related to respect, the valuing of status, honor, recognition and prestige	Apologize, class, courage, exclusive, notable	245	Lasswell
Strong	Words implying strength	Arose, attack, clout, enhance, prohibit, rampant	1,902	Harvard IV-4
SureLw	Words indicating a feeling of sureness, certainty and firmness	Absolute, bound, crucial, emphasis, fundamental, insistent, obvious, unlimited	175	Lasswell
Think	Words referring to the presence or absence of rational thought processes	Cognizant, esoteric, infer, morale, scrutinize, visionary	81	Harvard IV-4
Try	Words indicating activities taken to reach a goal, but not including words indicating that the goals have been achieved	Apply, endeavor, seek, strive, venture	70	Harvard IV-4
Undrst	Words indicating de-emphasis and caution	Accident, approximate, contingent, doubt, gradual, luck, nominal, speculate	319	Harvard IV-4
Vice	Words indicating an assessment of moral disapproval or misfortune	Acrimony, bizarre, capricious, cynical, misfortune, poverty, threat	685	Harvard IV-4
Virtue	Words indicating an assessment of moral approval or good fortune, especially from the perspective of middle-class society	Adaptable, beneficial, charisma, commitment, impressive, miracle, palatable, sincere, valuable	719	Harvard IV-4
Weak	Words implying weakness	Absent, afraid, anxiety, decline, fail, poor, unfortunate	755	Harvard IV-4
WlbPsyc	Words connoting the psychological aspects of well-being, including its absence	Anger, anxiety, bitter, calm, dread, furious, grief, happiness, mood, relieve, sad, terror, tragic	139	Lasswell
Yes	Words directly indicating agreement, including word senses "of course", "to say the least", "all right".	Agree, okay, sure, yeah, yes	20	Harvard IV-4

Figure 1a Daily Category Score Distributions (1991-2002): Same Scale

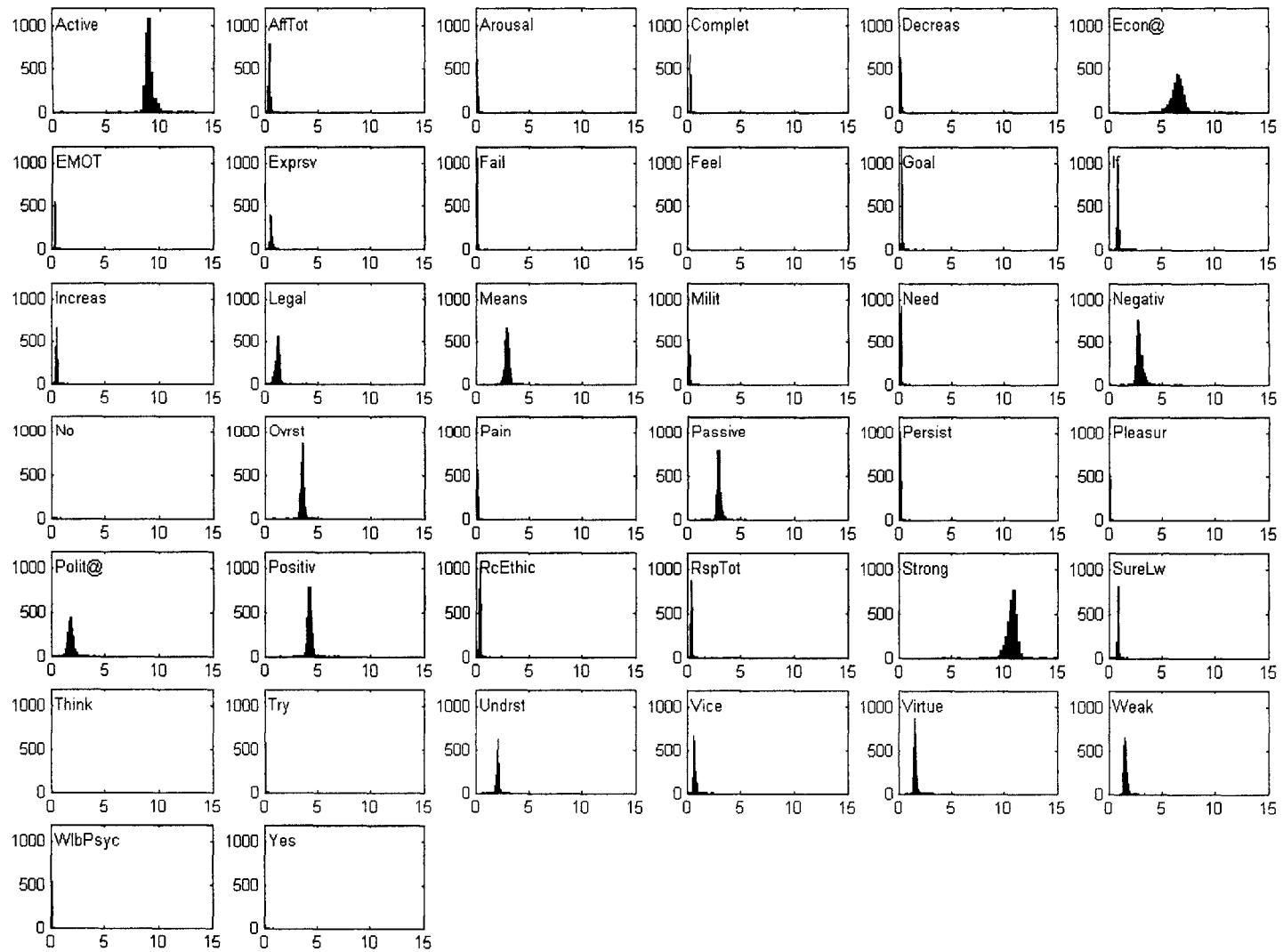


Figure 1b Daily Category Score Distributions (1991-2002): Individual Scales

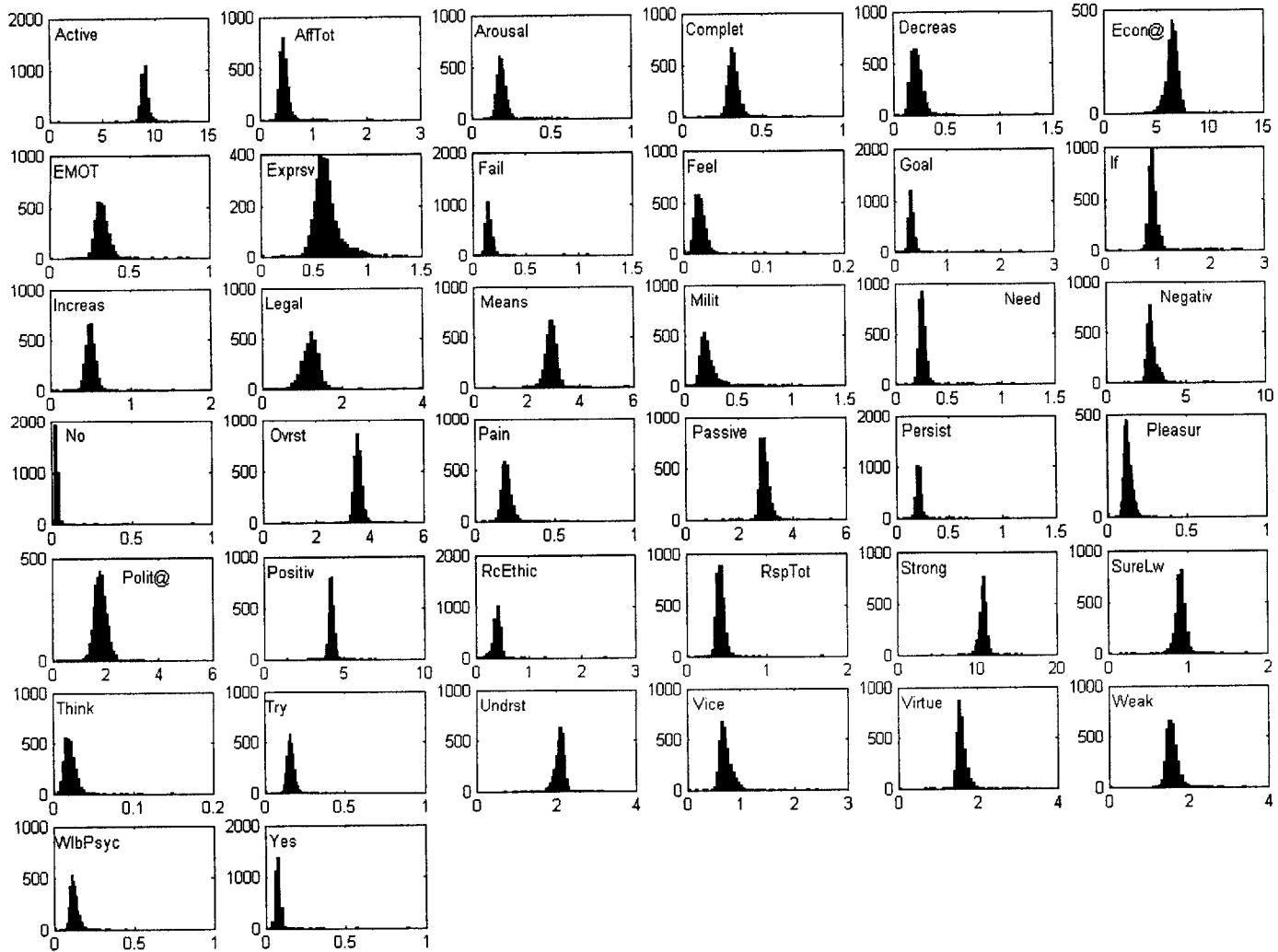


Table 2 Daily Category Score Statistical Properties (1991-2002)³

Score	Mean	Stdev	Skewness	Kurtosis	AutoCorr (1)	AutoCorr (2)	AutoCorr (3)	AutoCorr (4)	AutoCorr (5)	Box-Pierce (5)	Box-Pierce (5) P-value	Box-Pierce (20)	Box-Pierce (20) P-value	Min	5%	50%	95%	Max
Active	8.98	0.30	0.88	1.32	50.8%	41.9%	40.7%	51.3%	62.7%	3810.0	0.00	12962.7	0.00	8.20	8.56	8.95	9.59	10.22
AffTot	0.46	0.08	1.46	6.31	8.8%	9.3%	7.1%	9.1%	23.2%	250.4	0.00	632.6	0.00	0.23	0.36	0.45	0.60	1.22
Arousal	0.19	0.03	0.59	1.10	11.1%	10.3%	7.5%	10.7%	18.3%	223.6	0.00	700.7	0.00	0.12	0.15	0.19	0.23	0.34
Compleat	0.32	0.03	0.30	0.36	14.6%	7.7%	9.8%	13.8%	16.8%	255.7	0.00	632.8	0.00	0.19	0.27	0.32	0.37	0.44
Decreas	0.22	0.05	0.72	1.01	62.4%	55.8%	55.1%	57.2%	60.2%	5169.2	0.00	18032.6	0.00	0.11	0.15	0.21	0.31	0.51
Econ@	6.48	0.50	-0.57	0.88	43.6%	39.2%	39.0%	44.6%	56.0%	3073.9	0.00	10175.6	0.00	3.89	5.57	6.52	7.23	8.04
EMOT	0.33	0.04	0.59	1.04	25.7%	25.5%	22.2%	21.1%	27.2%	909.4	0.00	2719.6	0.00	0.22	0.27	0.33	0.39	0.55
Exprsv	0.62	0.12	1.48	3.85	11.5%	13.4%	11.2%	16.7%	48.1%	919.0	0.00	3091.8	0.00	0.35	0.47	0.60	0.86	1.35
Fail	0.15	0.03	1.16	2.90	36.0%	34.7%	35.8%	32.5%	35.3%	1841.7	0.00	6454.9	0.00	0.08	0.11	0.15	0.20	0.34
Feel	0.02	0.01	0.65	0.76	8.8%	7.7%	3.8%	7.3%	4.1%	57.9	0.00	144.0	0.00	0.01	0.01	0.02	0.03	0.05
Goal	0.32	0.05	0.63	0.68	33.4%	24.1%	22.6%	29.4%	36.4%	1343.8	0.00	2210.8	0.00	0.14	0.25	0.32	0.41	0.54
If	0.91	0.07	0.69	1.12	40.7%	39.3%	38.6%	37.7%	41.1%	2359.0	0.00	7933.5	0.00	0.68	0.81	0.90	1.04	1.30
Increas	0.50	0.06	0.16	0.33	23.7%	18.7%	16.8%	21.2%	28.8%	753.9	0.00	1985.6	0.00	0.24	0.41	0.49	0.59	0.72
Legal	1.21	0.18	-0.23	-0.04	56.7%	47.1%	46.7%	53.2%	58.1%	4182.7	0.00	14860.3	0.00	0.61	0.88	1.22	1.49	1.85
Means	2.91	0.19	-0.31	0.54	46.7%	33.6%	33.4%	45.4%	58.2%	3007.7	0.00	9920.5	0.00	1.86	2.58	2.93	3.21	3.65
Milit	0.22	0.08	2.70	13.11	60.1%	58.3%	55.5%	54.7%	51.6%	4786.9	0.00	12860.1	0.00	0.09	0.14	0.20	0.37	0.97
Need	0.25	0.03	0.43	0.60	18.5%	19.6%	18.3%	13.7%	20.4%	506.2	0.00	1697.1	0.00	0.17	0.21	0.25	0.30	0.38
Negativ	2.88	0.26	0.80	0.85	69.7%	66.9%	66.6%	66.7%	68.7%	6929.9	0.00	23668.1	0.00	2.15	2.53	2.84	3.41	4.31
No	0.03	0.02	31.58	1406.72	3.6%	-0.7%	0.7%	1.4%	4.4%	10.6	0.06	19.2	0.51	0.01	0.02	0.03	0.05	0.89
Ovrst	3.56	0.13	0.25	0.58	29.8%	14.4%	14.5%	25.7%	33.3%	936.9	0.00	2174.5	0.00	3.10	3.35	3.55	3.78	4.38
Pain	0.20	0.03	1.39	6.64	37.4%	36.1%	32.6%	27.8%	30.8%	1669.8	0.00	4710.0	0.00	0.13	0.16	0.20	0.25	0.51
Passive	2.96	0.16	1.01	1.95	52.1%	51.6%	49.8%	50.0%	51.6%	3944.4	0.00	14540.2	0.00	2.53	2.74	2.93	3.25	3.80
Persist	0.21	0.02	0.37	0.34	22.6%	20.3%	18.3%	21.3%	21.4%	659.8	0.00	2034.4	0.00	0.15	0.18	0.21	0.25	0.31
Pleasur	0.13	0.02	0.67	0.76	17.8%	19.9%	19.2%	18.2%	31.4%	730.3	0.00	2422.5	0.00	0.07	0.10	0.13	0.17	0.24
Polit@	1.81	0.23	0.59	2.70	42.4%	33.3%	32.3%	34.7%	39.7%	2055.0	0.00	4846.4	0.00	0.90	1.46	1.81	2.19	3.44
Positiv	4.23	0.17	0.23	0.34	34.4%	32.3%	29.3%	28.4%	31.0%	1481.2	0.00	4353.1	0.00	3.48	3.97	4.22	4.51	5.02
RcEthic	0.40	0.07	-0.58	0.88	60.9%	59.2%	57.7%	57.7%	60.3%	5293.2	0.00	19600.3	0.00	0.16	0.27	0.40	0.49	0.61
RspTot	0.42	0.05	0.55	0.81	18.0%	12.0%	9.5%	12.7%	20.1%	340.6	0.00	1014.2	0.00	0.27	0.35	0.42	0.51	0.64
Strong	10.78	0.40	-0.52	0.21	42.6%	23.6%	23.4%	42.3%	59.8%	2522.6	0.00	8780.9	0.00	9.46	10.01	10.82	11.37	11.87
SureLw	0.91	0.06	0.71	13.20	32.5%	25.4%	25.7%	30.7%	40.0%	1488.4	0.00	5358.7	0.00	0.69	0.81	0.91	1.00	1.73
Think	0.02	0.01	1.37	5.30	8.2%	8.2%	9.1%	7.7%	8.5%	106.2	0.00	283.9	0.00	0.01	0.01	0.02	0.03	0.09
Try	0.16	0.02	0.48	0.96	15.2%	7.3%	9.4%	15.6%	19.1%	298.5	0.00	796.4	0.00	0.10	0.13	0.16	0.20	0.29
Undrst	2.09	0.11	-0.67	0.71	51.8%	52.1%	51.0%	52.1%	61.8%	4407.9	0.00	16949.7	0.00	1.58	1.88	2.10	2.24	2.43
Vice	0.71	0.10	0.92	1.17	69.3%	66.7%	65.7%	63.3%	63.4%	6571.2	0.00	17162.9	0.00	0.49	0.58	0.69	0.91	1.32
Virtue	1.58	0.10	0.84	1.11	36.1%	34.1%	33.1%	34.7%	40.6%	1948.2	0.00	6694.3	0.00	1.33	1.44	1.57	1.78	2.08
Weak	1.58	0.13	0.51	0.39	58.3%	56.9%	56.1%	56.4%	56.7%	4864.7	0.00	17424.4	0.00	1.23	1.40	1.57	1.80	2.13
WlbPsc	0.13	0.03	1.56	5.88	36.3%	32.8%	30.9%	33.2%	37.9%	1793.7	0.00	5623.4	0.00	0.07	0.09	0.12	0.17	0.34
Yes	0.08	0.02	23.14	949.27	7.2%	6.3%	6.0%	2.6%	4.8%	48.0	0.00	144.6	0.00	0.04	0.06	0.08	0.10	0.90

³ Box-Pierce is the Ljung and Box corrected Box Pierce Q -statistic, $Q_m' \equiv T(T+2) \sum_{k=1}^m \frac{\rho^2(k)}{T-k}$, where T = sample series size, ρ = autocorrelation coefficient, k = sample series data, m = number of lags; formula from The Econometrics of Financial Markets by Lo, Campbell and MacKinlay, 1997, p. 47.

Table 3 Daily Category Score Correlations (1991-2002)

	Positiv	Negativ	Active	AffTot	Arousal	Comple	Decreas	Econ@	EMOT	Exprsv	Fail	Feel	Goal	If	Increas	Legal	Means	Milit	Need	No	Ovrst	Pain	Passive	Persist	Pleasur	Polit@	RcEthic	RspTot	Strong	SureLw	Think	Try	Undrst	Vice	Virtue	Weak	WibPsc	Yes	
Positiv	100%	16%	33%	30%	25%	25%	-1%	9%	20%	-5%	11%	9%	17%	28%	-1%	2%	1%	-3%	33%	0%	23%	13%	32%	9%	21%	7%	-2%	24%	20%	0%	0%	16%	0%	16%	51%	26%	9%	10%	
Negativ		100%	43%	7%	10%	6%	47%	5%	31%	-16%	55%	9%	9%	28%	-8%	6%	-5%	44%	11%	-1%	23%	46%	52%	21%	-1%	19%	4%	17%	22%	-19%	12%	15%	-12%	74%	10%	66%	26%	9%	
Active			100%	11%	16%	27%	16%	9%	16%	-13%	27%	7%	11%	30%	12%	-12%	8%	7%	32%	5%	18%	22%	41%	15%	7%	4%	-27%	18%	31%	-21%	9%	28%	-19%	39%	29%	42%	12%	8%	
AffTot				100%	41%	3%	-13%	-35%	22%	19%	3%	2%	-17%	18%	-22%	-18%	-27%	-3%	32%	8%	4%	19%	19%	-3%	32%	-17%	-12%	24%	-35%	10%	-2%	8%	3%	7%	25%	10%	31%	12%	
Arousal					100%	13%	3%	-16%	25%	8%	-5%	20%	13%	20%	-9%	-34%	-22%	-3%	11%	11%	21%	33%	33%	8%	28%	-16%	-34%	5%	-12%	13%	15%	8%	16%	20%	29%	20%	29%	-4%	
Comple						100%	4%	12%	5%	-12%	0%	5%	12%	12%	16%	-1%	7%	-5%	20%	2%	3%	5%	23%	10%	5%	1%	-11%	-3%	18%	-5%	17%	2%	3%	7%	16%	5%	1%		
Decreas							100%	23%	16%	-22%	26%	9%	16%	12%	-1%	0%	22%	20%	2%	-5%	23%	24%	34%	15%	-6%	23%	-6%	-9%	33%	-2%	0%	6%	4%	35%	0%	46%	4%	-3%	
Econ@								100%	-19%	-50%	5%	-5%	40%	14%	44%	17%	61%	-18%	-14%	-6%	7%	-11%	-4%	19%	-35%	-2%	14%	-19%	48%	-13%	-13%	-11%	17%	3%	-7%	29%	-40%	-7%	
EMOT									100%	8%	6%	8%	-8%	17%	-16%	-5%	-20%	17%	31%	0%	27%	61%	37%	16%	34%	11%	-12%	6%	-1%	10%	15%	15%	4%	24%	9%	23%	48%	-1%	
Exprsv										100%	-9%	-2%	-20%	-25%	-21%	-19%	-35%	-5%	-7%	6%	4%	-7%	-13%	-10%	35%	-17%	-6%	13%	-43%	11%	13%	9%	-11%	-18%	-1%	-28%	30%	1%	
Fail											100%	6%	5%	-4%	-15%	8%	-5%	23%	0%	3%	3%	7%	35%	6%	-6%	11%	20%	23%	12%	-22%	7%	12%	-18%	35%	2%	44%	8%	9%	
Feel												100%	15%	5%	-1%	-5%	-9%	-2%	1%	-2%	11%	8%	17%	-1%	9%	6%	-8%	7%	8%	2%	14%	8%	3%	11%	2%	13%	5%	-5%	
Goal													100%	19%	29%	-11%	24%	-7%	-19%	-4%	20%	1%	23%	1%	0%	1%	-9%	11%	36%	-17%	-2%	-17%	9%	19%	6%	32%	-14%	-19%	
If														100%	2%	-35%	-13%	3%	33%	0%	44%	35%	34%	38%	9%	-15%	-38%	-2%	4%	13%	7%	-9%	20%	40%	45%	50%	11%	1%	
Increas															100%	1%	31%	-15%	-11%	-9%	5%	-20%	-4%	9%	-10%	-2%	-4%	-7%	34%	-7%	-2%	-11%	15%	-9%	-6%	5%	-28%	-9%	
Legal																100%	38%	-1%	-14%	-11%	-34%	-19%	-17%	-24%	-32%	50%	77%	-4%	36%	-5%	-13%	-3%	15%	-12%	-47%	-18%	-27%	9%	
Means																	100%	-20%	-11%	-9%	-21%	-13%	-34%	19%	32%	-14%	49%	-4%	-14%	-14%	19%	-8%	-24%	7%	-39%	-1%			
Milit																		100%	3%	-4%	7%	19%	16%	11%	-4%	16%	2%	-4%	15%	-5%	1%	11%	-5%	24%	-1%	13%	22%	-1%	
Need																			100%	-4%	14%	17%	29%	20%	14%	-7%	-18%	3%	-13%	1%	-1%	15%	-9%	6%	27%	16%	14%	17%	
No																				100%	-5%	0%	11%	11%	-3%	-8%	-9%	10%	-12%	7%	0%	0%	-7%	4%	6%	5%	-5%	38%	
Ovrst																					100%	27%	24%	32%	16%	-9%	-29%	-4%	13%	43%	19%	0%	20%	43%	33%	32%	20%	-11%	
Pain																						100%	16%	19%	2%	-23%	-3%	-1%	2%	12%	10%	-3%	43%	16%	42%	45%	5%		
Passive																							100%	20%	21%	4%	-24%	17%	9%	-16%	12%	17%	-8%	43%	31%	59%	26%	-6%	
Persist																								100%	-1%	-13%	-22%	-2%	-1%	10%	15%	0%	16%	21%	24%	33%	11%	3%	
Pleasur																									100%	-17%	-22%	18%	-27%	11%	12%	9%	-5%	5%	20%	-1%	44%	6%	
Polit@																										100%	36%	-1%	46%	-4%	-8%	7%	2%	10%	-15%	3%	-10%	6%	
RcEthic																											100%	0%	21%	5%	-13%	-11%	19%	-11%	-45%	-18%	-29%	13%	
RspTot																												100%	-9%	-13%	2%	-7%	-18%	14%	13%	16%	6%	15%	
Strong																													100%	-6%	-8%	1%	17%	22%	-7%	19%	-21%	-10%	
SureLw																														100%	9%	-2%	51%	-5%	10%	-13%	8%	16%	
Think																														100%	9%	3%	10%	4%	9%	14%	-4%		
Try																															100%	-9%	6%	19%	7%	19%	-1%		
Undrst																																100%	2%	-5%	8%	-7%	-14%		
Vice																																	100%	24%	64%	25%	0%		
Virtue																																		100%	30%	25%	0%		
Weak																																				100%	12%	2%	
WibPsc																																						100%	-1%
Yes																																							100%

2.3 Market Variables

Since we're aiming for a macro-level analysis, we decided to complement our choice of the entire Wall Street Journal content with broad market indices variables. We included primarily three characteristics of the market variables to examine: return, volume, and volatility. A list of the market variables used is shown below⁴:

- S&P500 return
- S&P500 return square
- NYSE volume return
- CBOE VIX first difference
- 10-year US Treasury return
- DEM(EURO)/USD average daily exchange rate return
- YEN/USD average daily exchange rate return
- West Texas Intermediate Oil Price (US\$/Barrel) return
- Gold Bullion Price-New York (US\$/Ounce) return
- CBOE Put/Call ratio
- NYSE Advance/Decline ratio

Market returns and return squares are calculated as follows. We used return squares as a proxy for daily market volatility. For exchange rate return calculations, the close value is replaced by daily average price. For volume return calculation, the close value is replaced by total daily volume:

$$Return_t = \frac{Close_t - Close_{t-1}}{Close_{t-1}}$$

$$ReturnSquare_t = (Return_t)^2$$

First difference calculations are as follows:

$$FirstDifference_t = Close_t - Close_{t-1}$$

Statistical properties of the daily market variables chosen are summarized in Table 4.

⁴ S&P500, oil, and gold data from Global Financial Database. CBOE put/call and NYSE advance/decline data from topline-charts.com. NYSE volume from NYSE.com. VIX from Yahoo Finance. 10-year US Treasury from Ryan Labs. Exchange rates from OANDA.com

Table 4 Daily Market Variables Statistical Properties (1991-2002)

Market Variable	Mean	Stddev	Skew- ness	Kurtosis	Auto- Corr (1)	Auto- Corr (2)	Auto- Corr (3)	Auto- Corr (4)	Auto- Corr (5)	Box- Pierce (5)	BP (5) P-value	Box- Pierce (20)	BP (20) P-value	Min	5%	50%	95%	Max
S&P500 Return	3.9E-04	1.1E-02	0.0	3.9	0.1%	-2.7%	-3.7%	0.2%	-3.4%	10.0	7.63E-02	32.3	4.02E-02	-6.9E-02	-1.7E-02	2.8E-04	1.7E-02	5.7E-02
S&P500 Return Square	1.1E-04	2.7E-04	7.5	85.4	20.8%	19.3%	20.1%	14.8%	19.6%	549.3	0.00	1282.9	0.00	0.0E+00	1.8E-07	2.8E-05	4.8E-04	4.7E-03
NYSE Volume Return	0.02	0.22	3.1	24.4	-26.6%	-12.1%	-4.3%	-3.1%	11.9%	309.5	0.00	427.9	0.00	-0.75	-0.25	0.00	0.30	2.20
CBOE VIX 1st Diff	2.1E-03	1.46	0.6	8.7	-10.3%	-6.2%	-6.5%	0.5%	-6.6%	69.6	0.00	119.5	0.00	-9.50	-2.14	-0.02	2.29	13.77
10 Yr US Treasury Index Return	1.5E-05	6.2E-03	1.0	68.2	4.9%	-1.4%	-6.1%	-3.3%	-2.0%	24.5	1.72E-04	44.1	1.45E-03	-7.3E-02	-7.4E-03	0.0E+00	6.9E-03	9.9E-02
DEM(EURO)/USD Return	6.7E-05	5.4E-03	0.1	3.5	1.4%	2.7%	-1.6%	-0.5%	1.0%	5.8	0.33	14.8	0.79	-3.2E-02	-8.8E-03	0.0E+00	8.7E-03	3.4E-02
YEN/USD Return	-1.4E-05	5.5E-03	-0.7	7.8	10.8%	4.7%	-2.7%	-1.7%	-3.0%	68.8	0.00	89.1	0.00	-6.1E-02	-8.7E-03	0.0E+00	8.8E-03	2.8E-02
Oil Return	3.5E-04	0.02	-0.8	17.5	-1.0%	-5.7%	-8.5%	-1.1%	0.8%	32.5	5.00E-06	46.2	7.47E-04	-0.33	-3.6E-02	5.1E-04	3.5E-02	0.16
Gold Return	-3.2E-05	7.6E-03	0.6	15.2	1.3%	-2.2%	-1.0%	2.5%	6.5%	17.1	4.31E-03	40.4	4.49E-03	-7.4E-02	-1.1E-02	-1.6E-04	1.2E-02	9.4E-02
CBOE Put/Call	0.72	0.16	0.6	0.7	63.1%	50.3%	49.1%	45.0%	41.2%	3831.0	0.00	10038.5	0.00	0.15	0.48	0.70	1.00	1.56
NYSE Advance/Decline	1.15	0.59	1.7	7.3	18.1%	2.7%	2.6%	3.5%	0.9%	107.6	0.00	138.7	0.00	0.06	0.40	1.05	2.22	6.83

2.4 Event Study

To help us better understand the impact of important events on category scores, we compiled and categorized a list of important event dates. Category score statistics are then computed before and after these important dates and compared to their entire period statistics. With this data, the goal is to observe how category score values change before, during, and after the occurrence of major events. The events collected are independent of any one news source. Major event types of the events collected and a brief explanation of each are given below⁵:

- **Accidents:** Plane crashes, nuclear accidents, etc.
- **Business - Bankruptcy & Downsizing:** Bankruptcies and layoffs
- **Business - M&A:** Mergers and acquisitions
- **Business - Other:** Business events that do not fall under M&A and bankruptcy/downsizing
- **Financial:** S&P record drops, credit spread movements, etc.
- **Macroeconomic:** Fed rates changes, GDP, unemployment, etc.
- **Military:** Wars, weapons inspections, etc.
- **Natural Disaster:** Earthquakes, floods, fires, etc.
- **Political - International:** Treaties, elections, assassinations, etc. outside of the US
- **Political - US:** Treaties, elections, assassinations, etc. within the US
- **Terrorism:** 9/11, Oklahoma bombing, etc.

For financial events, we define significant rises, drops, and movements in major indices as different events. For example, a significant rise in the S&P500 is defined as a day t on which the return $Return_t$ is greater than some positive constant multiple of the entire period standard deviation of the return, $n \cdot \sigma_{Return}$. Likewise, a significant drop in the S&P500 is defined as a day t on which the return $Return_t$ is less than some negative constant multiple of the entire period standard deviation of the return, $-n \cdot \sigma_{Return}$. A significant movement is simply a significant rise or drop. The constant multiple n is adjusted so as to produce sufficient number of events for each financial event category, usually between 50 and 200 events for the entire period of 1991 to 2002. We fixed $n = 2$ in our analysis to stay in this range.

$$Rise_t : Return_t > n \cdot \sigma_{Return}$$

$$Drop_t : Return_t < -n \cdot \sigma_{Return}$$

⁵ Events collected from a variety of sources. Most events are from the following sources: 1) The World Almanac and Book of Facts, 2) World Political Almanac 4th Edition, 3) CIA World Fact Book, 4) Wikipedia, 5) Information Please Almanac, 6) Life Magazine Year in Pictures, 7) Boston Globe's Year in Review, 8) Wall Street Journal, 9) Economic Perspectives, Federal Reserve Bank of Chicago, 10) Economic Indicators, State of California Department of Finance

$$Move_t : |Return_t| > n \cdot \sigma_{Return}$$

Financial markets used to compute financial variables include Credit Spread, Nasdaq 100, S&P500, VIX, DEM(Euro)/USD, and Yen/USD⁶.

After events of all event types are compiled, we deleted events of the same type whose dates are spaced too close together. The reason for doing so is to eliminate overlap in the events that would skew event study results. When we examine the impact of important events on category scores around event occurrence dates, we set a window for the number of days to examine the category scores before and after the events occur. If events of the same type are spaced too close together, the effects of an earlier event may undesirably propagate into the window that we're using to examine the effects of the next event. For our analysis, we decided on a 5 day window (5 days before and 5 days after an event occurs). Therefore, we eliminated events of the same type that are spaced within 10 days apart. More specifically, we kept the first event within any 10 day period and eliminated any and all subsequent events of the same type that occur within the next 9 days.

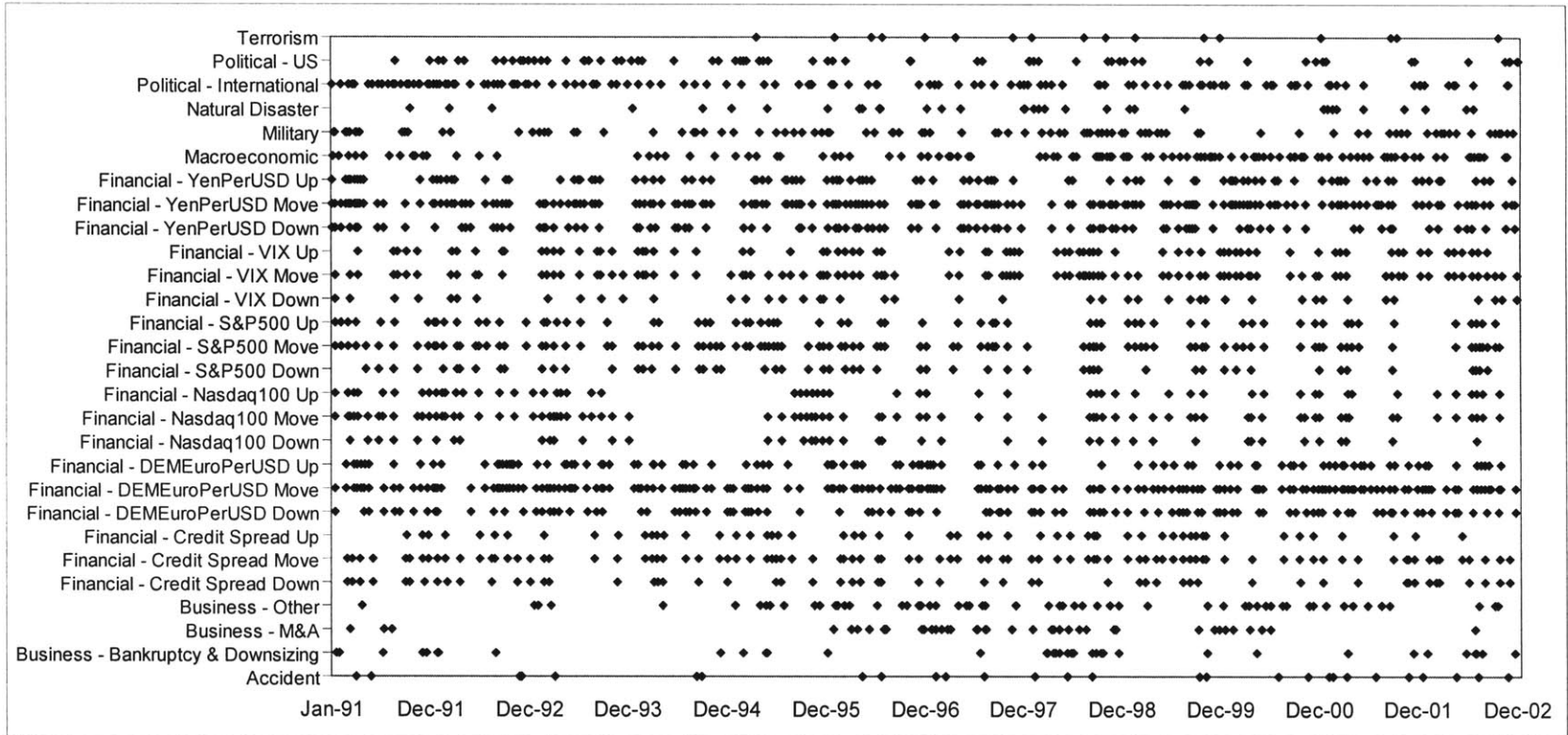
Table 5 shows the number of events for each year across each event category after we compiled the events list and eliminated events of the same type whose dates are spaced too close together. Figure 2 is a plot of the events across the time period 1991-2002, where a point represents an event of the type on that date. Note that on certain extreme days like the days following 9/11/01, there are events in multiple categories such as terrorist, military, financial, etc.

⁶ Credit Spread = KDP High Yield - US Treasury 10Yr. Credit Spread data from Bloomberg, Nasdaq 100 data from Yahoo! Finance, all other data from same sources as market variables

Table 5 Important Event Counts by Event Type and by Year (1991-2002)

Event Type	Event Count												Total
	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	
Accident	2	2	1	2	0	2	3	3	2	2	5	4	28
Business - Bankruptcy & Downsizing	5	3	0	1	3	1	1	14	1	1	2	6	38
Business - M&A	3	0	0	0	0	9	10	11	3	5	0	1	42
Business - Other	1	0	3	1	7	9	8	8	3	9	6	3	58
Financial - Credit Spread Down	7	5	4	5	3	6	3	3	6	2	5	7	56
Financial - Credit Spread Move	8	8	5	8	8	8	7	8	11	5	7	8	91
Financial - Credit Spread Up	3	4	3	6	6	5	5	6	9	4	3	1	55
Financial - DEMEuroPerUSD Down	8	6	9	9	8	8	6	10	11	11	9	9	104
Financial - DEMEuroPerUSD Move	13	15	15	16	10	18	13	11	15	15	18	16	175
Financial - DEMEuroPerUSD Up	8	11	10	9	5	13	8	3	7	8	13	8	103
Financial - Nasdaq100 Down	5	3	5	1	6	5	2	4	3	5	3	1	43
Financial - Nasdaq100 Move	12	8	11	1	8	6	4	6	6	7	5	6	80
Financial - Nasdaq100 Up	8	8	7	0	6	3	3	4	4	5	3	6	57
Financial - S&P500 Down	4	6	4	8	5	7	3	5	3	4	3	4	56
Financial - S&P500 Move	9	11	7	10	12	10	7	6	9	8	6	8	103
Financial - S&P500 Up	7	8	6	5	8	5	5	3	6	5	5	5	68
Financial - VIX Down	4	3	4	1	7	3	2	2	6	4	4	4	44
Financial - VIX Move	7	5	9	6	10	9	9	11	9	11	8	10	104
Financial - VIX Up	5	5	8	6	4	7	8	9	5	8	7	7	79
Financial - YenPerUSD Down	9	8	8	9	6	12	10	8	10	7	5	8	100
Financial - YenPerUSD Move	15	15	13	14	13	17	15	10	15	16	14	13	170
Financial - YenPerUSD Up	9	9	7	8	9	11	8	3	6	11	10	8	99
Macroeconomic	11	3	0	6	7	6	7	14	18	16	14	11	113
Military	11	3	8	6	8	9	6	12	9	2	7	14	95
Natural Disaster	1	2	0	2	2	5	4	7	3	0	6	3	35
Political - International	20	19	12	6	8	8	11	8	14	11	7	8	132
Political - US	1	10	11	6	8	5	2	8	5	3	4	6	69
Terrorism	0	0	0	0	1	4	2	5	3	1	2	1	19
Total	196	180	170	152	178	211	172	202	202	186	181	186	2216

Figure 2 Important Events Plot by Event Type (1991-2002)

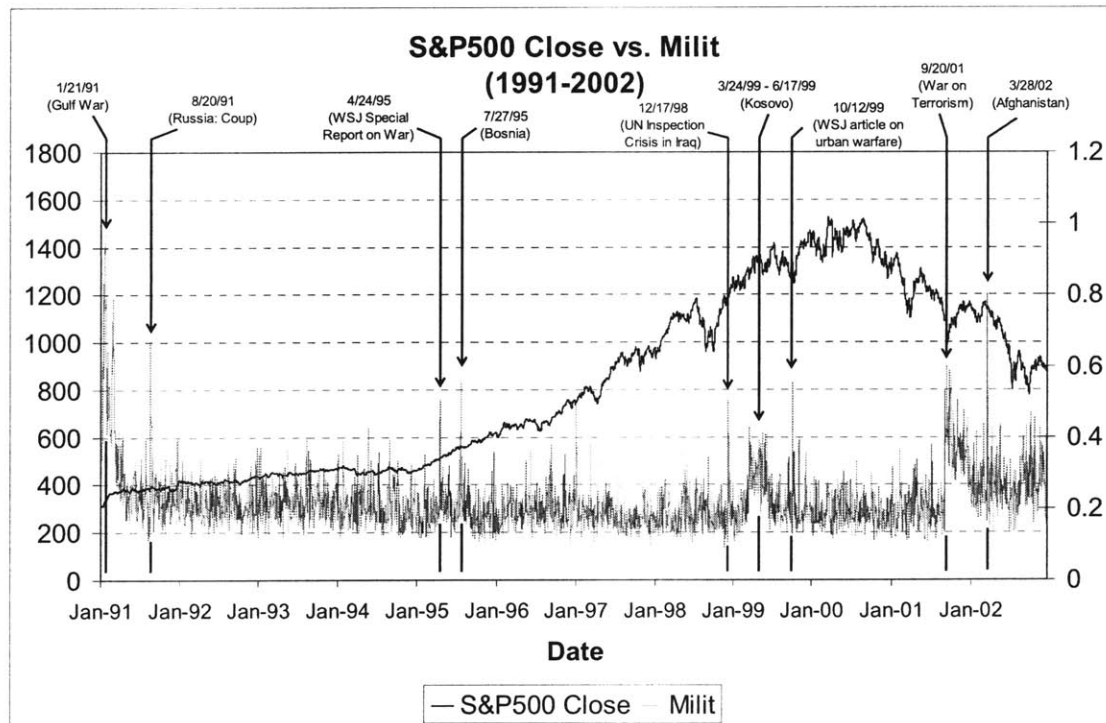


3. Methodology

3.1 Validity of GI-Extracted Content

Finding meaningful relationships between category content extracted by GI and market values rely on a few assumptions. First, we must assume that GI is extracting meaningful information from the Wall Street Journal. To test this assumption, we examined time series plots of subject content category scores such as Milit (Military), Polit@ (Political), and Econ@ (Economics) and investigate whether any important events in the respective categories occurred on dates where the category exhibited significantly high values. Figure 3 shows a time series plot of S&P500 closing value and the Milit category score from 1991 to 2002 annotated with important military events. As the plot shows, during times of important military events, there is a dramatic increase in the Milit category score. The results from this simple analysis provide some reassurance that GI is extracting meaningful content.

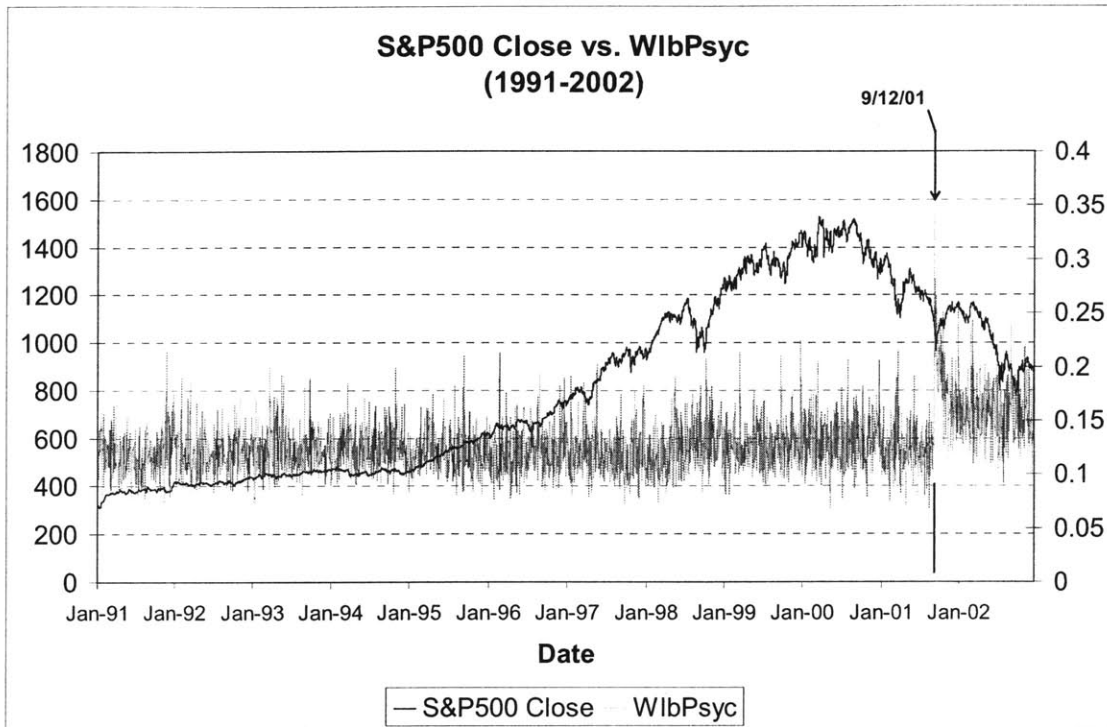
Figure 3 Time Series Plot of the Milit Category Score (1991-2002)



The assumption we must make going forward is that GI can extract emotional-type content with categories like WlbPsyc (Psychological Well-being) in the same fashion and as successfully as subject-type content with categories like Milit. Figure 4 shows a time

series plot of the WlbPsyc category score from 1991 to 2002. Note that on 9/12/01 there is a huge spike in the WlbPsyc score indicating that it is picking up relevant content and capturing the extreme psychological state of that time.

Figure 4 Time Series Plot of WlbPsyc Category Score (1991-2002)



3.2 Definition of Dimensions

The bulk of our analysis involves the discovery of relationships between category score and market variable time series. We now define the dimensions used when relating one series to another. There are three dimensions we incorporated into our analysis.

3.2.1 Dimension 1: Daily and Weekly

The first dimension deals with the incremental time period we're looking at the category and market values. We decided to look at the values on a daily and weekly basis. We did not include the monthly case as there would be only 12 data points per year which is not sufficient to draw statistically significant conclusions from correlations and regressions for yearly periods. The daily category scores of category i , $S_{i,daily}$, is calculated as the daily number of words that fall under that category, $WC_{i,daily}$, divided by the daily total number of words, TWC_{daily} :

$$S_{i,daily} = \frac{WC_{i,daily}}{TWC_{daily}}$$

We now explain the weekly category score and market value calculations. The weekly category score for a particular category i for week j , $S_{i,week_j}$, is computed as the sum of the number of daily word count in category i , $WC_{i,daily}$, over week j divided by the sum of the number of total daily word counts, TWC_{daily} , over week j :

$$S_{i,week_j} = \frac{\sum_{week_j} WC_{i,daily}}{\sum_{week_j} TWC_{daily}}$$

Weekly market returns for S&P500, 10-year treasury, exchange rates, gold, and oil market variables are computed as the geometric average of daily returns:

$$Return_{week_j} = [\prod_{week_j} (1 + Return_{daily})] - 1$$

The weekly return square is then just the square of the weekly return:

$$ReturnSquare_{week_j} = (Return_{week_j})^2$$

The weekly first difference is computed as the difference between this Friday's closing value and last Friday's closing value:

$$FirstDifference_{week_j} = Close_{Friday,week_j} - Close_{Friday,week_{j-1}}$$

The weekly volume is computed as the sum of the daily volumes in the week:

$$Volume_{week_j} = \sum_{week_j} Volume_{daily}$$

Volume return is computed as the total volume of this week minus the total volume of last week, divided by the total volume of last week:

$$VolumeReturn_{week_j} = \frac{Volume_{week_j} - Volume_{week_{j-1}}}{Volume_{week_{j-1}}}$$

For weekly volume return computations where the number of market days in one week is not equal to the number of market days in the previous week, we attempt to project the volume for the week with fewer than five market days. The projected volume for the shorter week with fewer market days is equal to the original weekly volume multiplied by the number of days in the normal week divided by the number of days in the week with fewer market days.

$$Volume_{projected, week_{short}} = Volume_{week_{short}} \cdot \frac{NumDays_{week_{long}}}{NumDays_{week_{short}}}$$

For weekly ratios such as the NYSE advance decline ratio, we simply sum the number of advances in the week and divide by the sum of the number of declines in the same week. The same calculation is performed for CBOE put call ratio.

$$NYSEAdvanceDecline_{week_j} = \frac{\sum_{week_j} NYSEAdvance_{daily}}{\sum_{week_j} NYSEDecline_{daily}}$$

$$CBOEPutCall_{week_j} = \frac{\sum_{week_j} CBOEPut_{daily}}{\sum_{week_j} CBOECall_{daily}}$$

3.2.2 Dimension 2: Leads and Lags

The second dimension addresses different combinations of market variables leading or lagging the category scores and vice versa. In conjunction with correlations and regressions, this dimension is used to detect causality. In other words, if we lag the category score time series relative to the market variable time series, the results of the correlations and regressions can help to answer questions such as “do emotions influence the market, or vice versa, or both?” and “if so, to what extent?” We naturally have to incorporate the first dimension of daily or weekly values when we deal with the lag dimension. For example, if we are dealing with weekly values, lagging the category scores relative to market scores by 1 unit is effectively lagging by 7 days.

3.2.3 Dimension 3: Yearly and Entire Period

The third dimension is simply the time period (the length of the time series) over which we run our analysis. We decided to run our analysis over individual years and over the entire period of the data set from January 2, 1991 to December 31, 2002. To give an example of how the three dimensions defined manifest themselves in our analysis, consider a correlation between weekly category and market values over the year of 1991 where category scores lag market values by 1 week.

3.3 Correlations, Regressions

To discover quantitative relationships between the category scores and market values, we used correlations and regressions.

We used the Pearson correlation coefficient in conjunction with the two-tailed T-test to help us determine the strength and statistical significance of the correlations. We used R-Square and the associated F-value and T-value of the parameter estimates as an indicator of the significance of regression results. We varied the dimensions mentioned in the previous section when running correlation and regression analysis. For example, lagged regressions can help us determine whether any of the category scores have predictive power for a particular market variable.

Missing data points are ignored in the correlation and regression analysis. For example, suppose we're correlating lagged market values against category scores such that we encounter a Friday market value matched with a nonexistent weekend category score. In this case, we ignore the Friday data point. It would be incorrect to pair the Friday market value with a Monday category score because we work with daily emotional states, which means Friday emotions are likely to dissipate by Monday. Likewise, weekend emotional states are reflected in Monday's markets and since we don't have weekend Wall Street Journal data in our analysis, it would be incorrect to pair Friday category scores with Monday market values.

3.4 Event Study

For each event category, we have a list of dates for all events in that category. With this information, we examine the statistical properties of the category scores up to n days before and after each event type happens by computing the statistics across all the events in the event type. The number n represents the width of the window or time span we're examining the category score statistics around major event occurrences. We've chosen $n = 5$ for our event study analysis.

So more formally, let $t_{i,k}$ be the date on which an event i of event type k happens. Let $S_{j,t_{i,k}}$ be the category score of category j on the date of the event, or date closest to the event date on which a category score is available. Then $S_{j,t_{i+1,k}}$ is the score on the date of another event $i+1$ of event type k . It follows that $S_{j,t_{i,k}+1}$ is the score one day after event i happens. Similarly, $S_{j,t_{i,k}-1}$ is the score one day before event i happens. If these scores are not available, as in the case of weekends for example, we use the next or last available score after the closest date to the date of the event that we're processing. A summary statistic, say average, that examines the impact of m events of event type k on category score S_j on the n^{th} day before or after each event happens can be expressed as follows:

$$EventAverage(S_j, k, n) = Average(\{S_{j,t_1,k+n}, S_{j,t_2,k+n}, \dots, S_{j,t_m,k+n}\}) = \frac{\sum_{i=1}^m S_{j,t_i,k+n}}{m}$$

The same event average statistics can be computed for market variables to investigate the impact of important events on the market indices. The category score S_j in the formula above is then replaced by R_j for the return of market variable j , for example.

We can also compute standard deviations of the category score S_j on the n^{th} day before or after all m events of type k happen:

$$EventStdev(S_j, k, n) = StandardDeviation(\{S_{j,t_1,k+n}, S_{j,t_2,k+n}, \dots, S_{j,t_m,k+n}\})$$

We can then define the normalized average category score $\overline{EventAverage(S_j, k, n)}$ for a particular event type k on the n^{th} day before or after the events happen as the event average score $EventAverage(S_j, k, n)$ divided by the entire period score average for that category, $\overline{S_j}$:

$$\overline{EventAverage(S_j, k, n)} = \frac{EventAverage(S_j, k, n)}{\overline{S_j}}$$

The normalized category score standard deviation is then:

$$\overline{EventStdev(S_j, k, n)} = \frac{EventStdev(S_j, k, n)}{(\overline{S_j})^2}$$

The normalized average and standard deviation tell us how the average and standard deviation statistics of the subject scores around event dates compare to the same statistics over the entire period. Next, to find excess the average, we subtract one from the normalized average:

$$EventExcessAverage(S_j, k, n) = \overline{EventAverage(S_j, k, n)} - 1$$

This excess average number tells us how much the average category score before or after events of a certain type deviates from the entire period average.

4. Results

We present the results of correlation and regression analysis below. Since weekly results do not show significant improvement over daily results, we will focus on daily results only. We also present below the impact of outlier data points on daily correlations as well as the results of event studies.

4.1 Correlation Results

Figure 5 shows the daily no lag correlations between each market variable and the top five most strongly correlated category scores. Correlations are shown for each year and for the entire period of 1991-2002. The top five most strongly correlated categories are the five categories with the highest sum square of correlations with the given market variable for each year. In other words, if c_i is the correlation coefficient between the category score and the market variable for year i , then the sum square, $c_{i,sum}$, is computed as:

$$c_{i,sum} = \sum_{i=1991}^{2002} (c_i)^2$$

Figure 6 and Figure 7 show the same correlation results for category scores lagging markets by 1 day and markets lagging category scores by 1 day, respectively. For yearly correlations, correlation strength of about 13% corresponds to a P-value of about 0.05. For the entire period of 1991-2002, correlation strength of about 9% corresponds to a P-value of about 0.05.

Figure 5 Top 5 Most Strongly Correlated Categories: Daily, No Lag

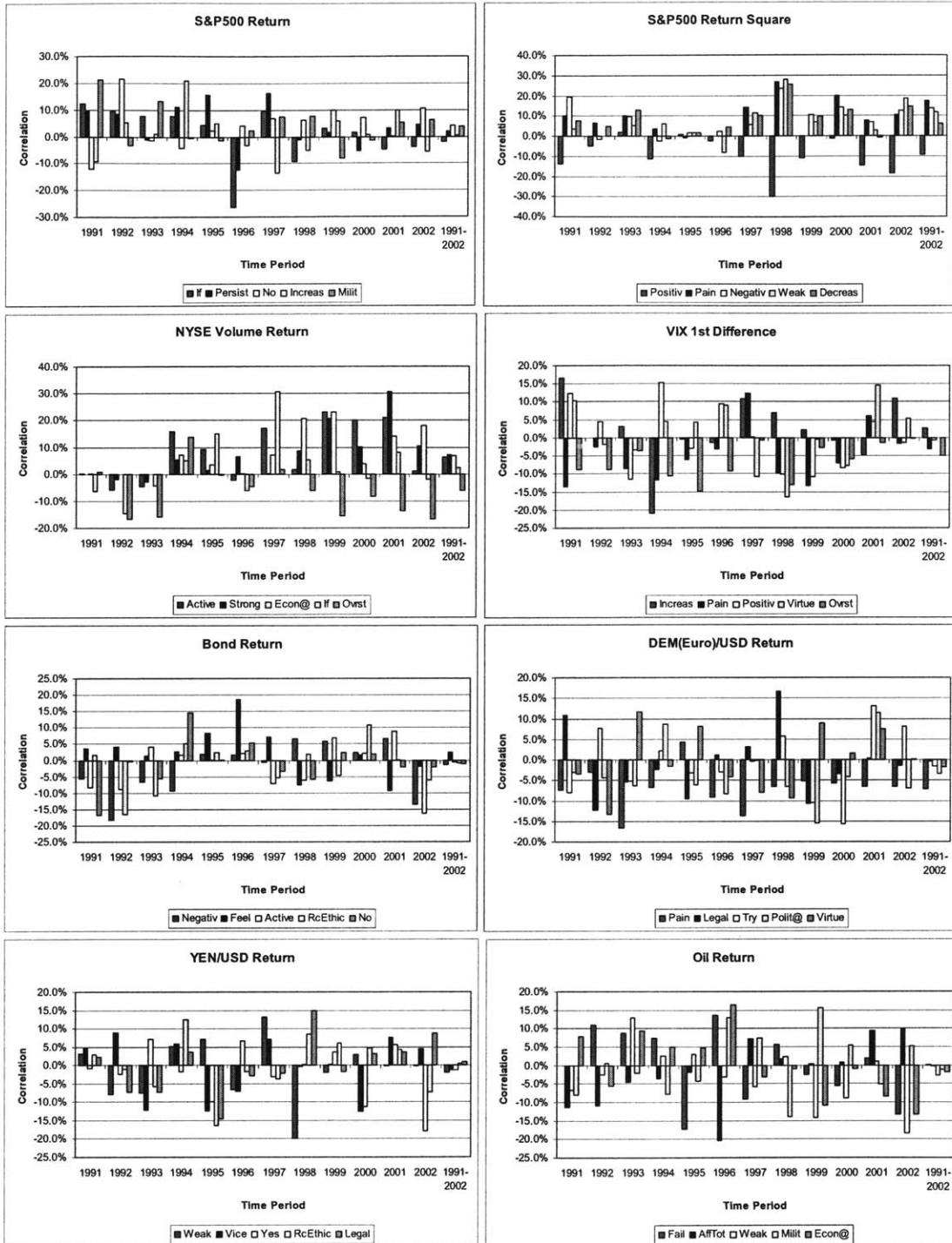


Figure 5 (Con't) Top 5 Most Strongly Correlated Categories: Daily, No Lag

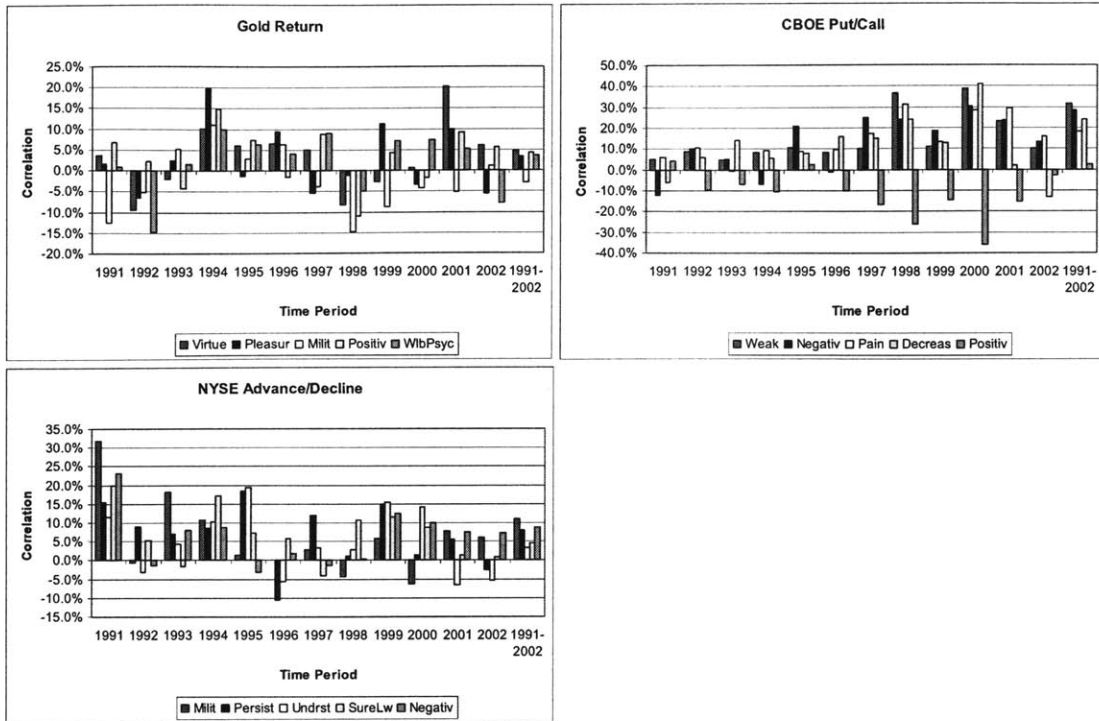


Figure 6 Top 5 Most Strongly Correlated Categories: Daily, Categories Lag Markets by 1 Day

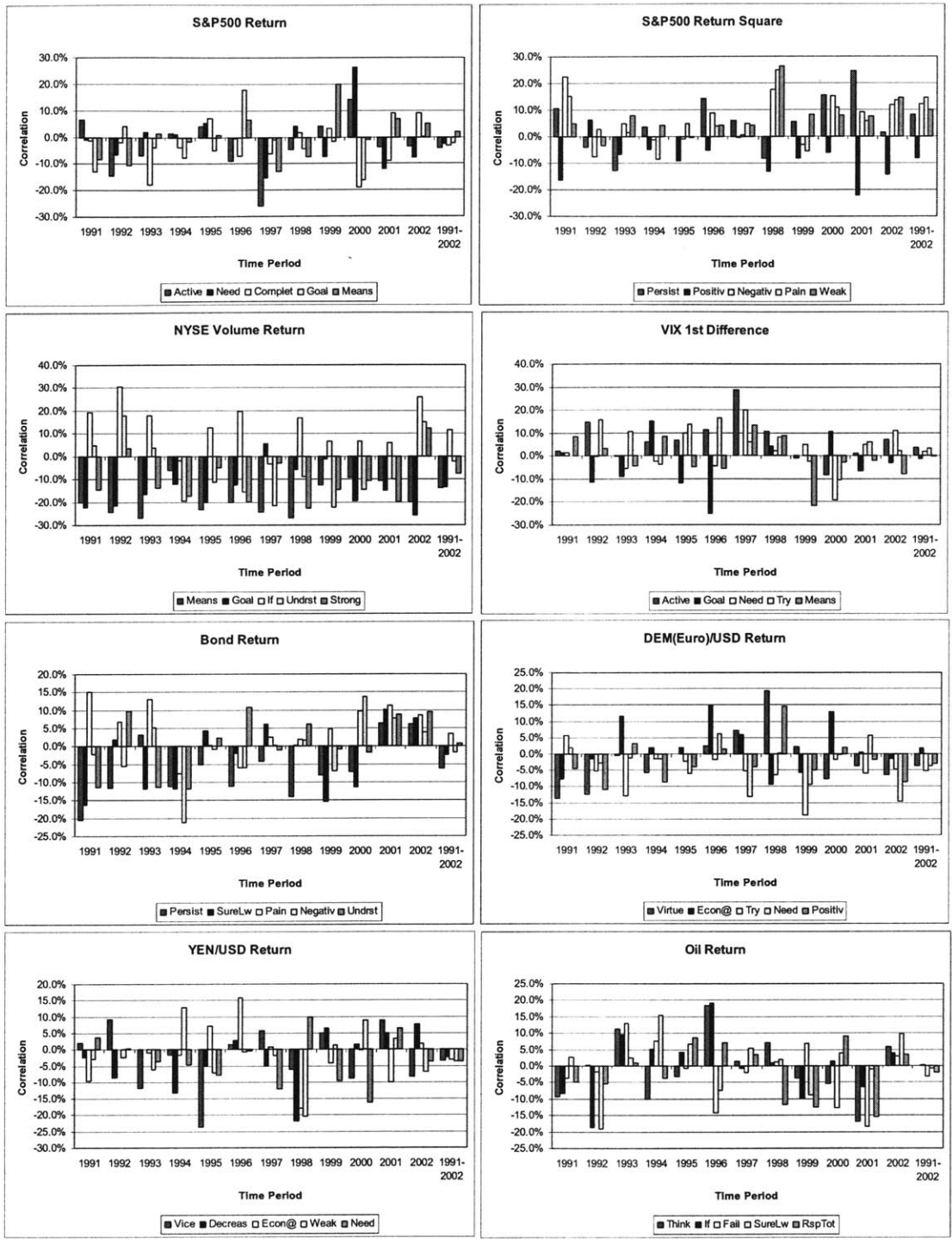


Figure 6 (Con't) Top 5 Most Strongly Correlated Categories: Daily, Categories Lag Markets by 1 Day

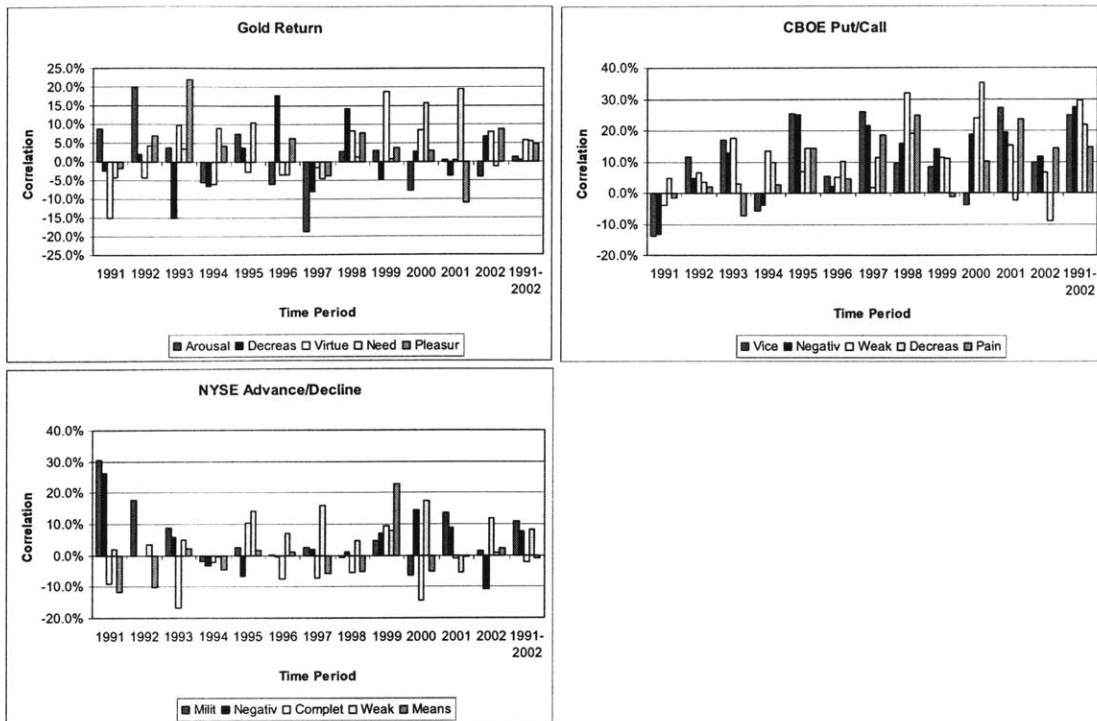


Figure 7 Top 5 Most Strongly Correlated Categories: Daily, Markets Lag Categories by 1 Day

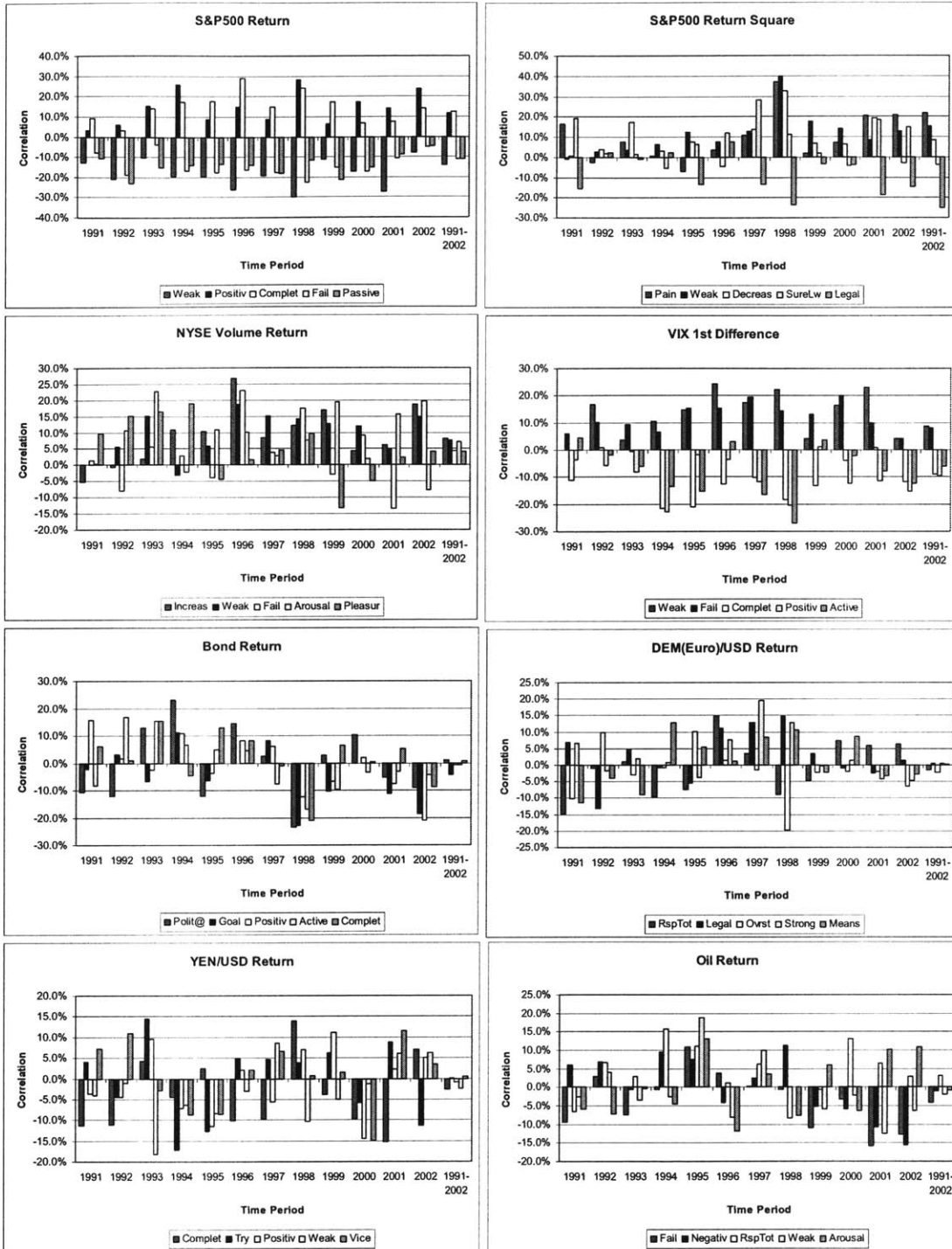
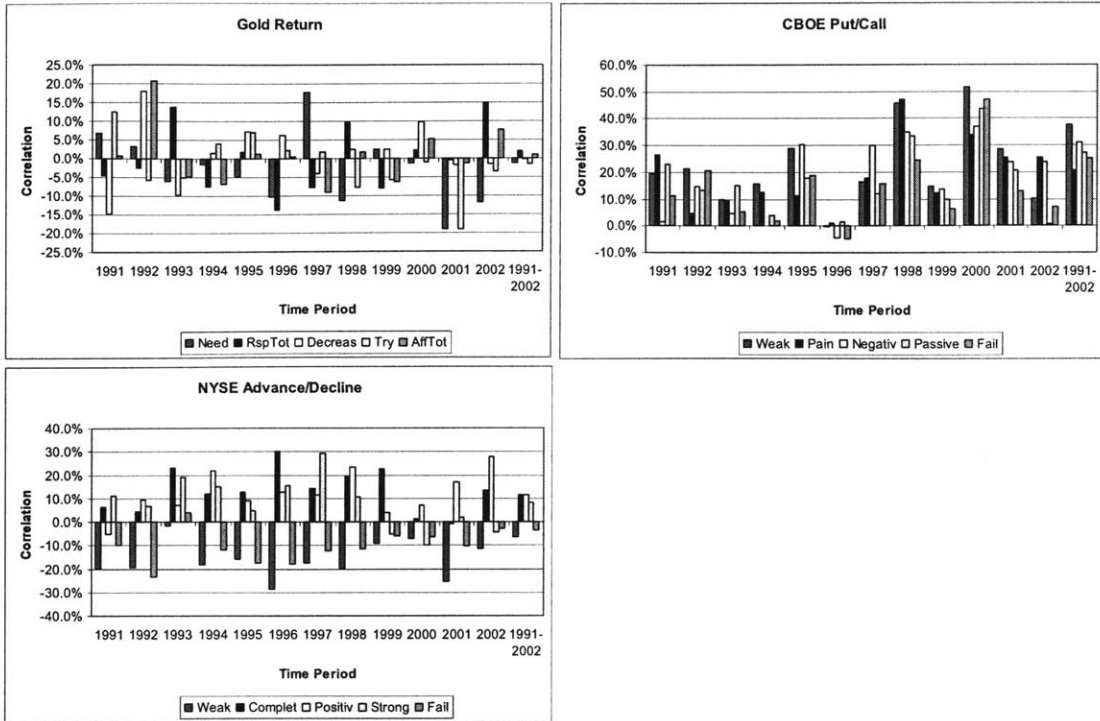


Figure 7 (Con't) Top 5 Most Strongly Correlated Categories: Daily, Markets Lag Categories by 1 Day



4.2 Regression Models

Below we present the daily stepwise linear regression models for Market = Categories, where the market variable value is the dependent variable and the category scores are the independent variables. Regressions were run over each individual year and for the entire period of 1991-2002. We've selected the 1-variable, 3-variable, and 5-variable models. Significance test statistics, namely F-value for R-Square and T-value for parameter estimates, are given in parenthesis. Tables 6.1-6.11 show the no lag case. Tables 7.1-7.11 show the case where category scores lag markets by one day.

Table 6.1 Daily Stepwise Linear Regressions: SP500Return = Categories, No Lag

SP500Return		1-Variable			3-Variable					5-Variable							
Period	Total Size	R-Square	Intercept	Milit	R-Square	Intercept	Fail	Milit	No	R-Square	Intercept	Fail	Increas	Milit	No	SureLw	
1991	252	4%	-0.0028 (11.76)	0.0125 (3.43)	9%	0.0095 (7.77)	-0.0450 (2.13)	0.0138 (2.03)	-0.1604 (3.83)	11%	0.0009 (5.96)	-0.0444 (0.06)	-0.0216 (-2.01)	0.0110 (1.96)	-0.1781 (2.80)	0.0223 (3.06)	(1.65)
1992	253	5%	-0.0052 (12.31)	0.1770 (3.51)	8%	-0.0329 (7.33)	0.1889 (-2.87)	0.0074 (3.78)	0.0410 (1.95)	10%	-0.0140 (5.44)	-0.0029 (-0.68)	0.0125 (1.45)	0.1893 (1.66)	0.0081 (3.81)	0.0429 (2.13)	(2.07)
1993	252	2%	-0.0029 (4.39)	0.0147 (1.87)	4%	0.0090 (3.65)	-0.0232 (1.76)	-0.0176 (-1.88)	0.0125 (1.79)	5%	-0.0036 (2.75)	-0.0241 (-0.27)	-0.0207 (-1.91)	0.0120 (2.14)	-0.0026 (1.71)	0.0017 (-1.35)	(1.27)
1994	251	4%	-0.0122 (11.31)	0.0233 (3.36)	8%	-0.0380 (6.99)	0.0259 (-4.01)	0.0306 (3.76)	0.0180 (1.87)	9%	-0.0386 (4.93)	0.0264 (-3.92)	0.0309 (3.84)	0.0187 (1.88)	-0.0316 (1.17)	0.0186 (-1.50)	(2.16)
1995	251	3%	-0.0231 (7.33)	0.0114 (2.58)	7%	-0.0334 (6.15)	0.0380 (-3.54)	0.0340 (2.54)	0.0104 (2.04)	9%	-0.0301 (4.97)	-0.0040 (-2.65)	0.0415 (-1.89)	0.0302 (2.76)	0.0029 (1.80)	0.0118 (1.79)	(2.84)
1996	253	7%	0.0339 (18.90)	-0.0379 (4.44)	10%	0.0517 (8.89)	0.1354 (4.14)	-0.0343 (1.98)	-0.0112 (-1.92)	13%	0.0403 (7.31)	0.1611 (2.95)	-0.0373 (2.37)	-0.0472 (-4.10)	0.0261 (-2.04)	-0.0133 (2.45)	(-2.26)
1997	252	3%	-0.0182 (6.92)	0.0936 (2.47)	8%	0.0122 (6.88)	-0.0412 (1.10)	-0.0736 (-3.16)	0.1175 (2.52)	11%	0.0095 (5.93)	0.0647 (2.33)	-0.2096 (-1.97)	-0.0462 (-3.55)	-0.0862 (-2.94)	0.1192 (3.36)	(3.36)
1998	251	2%	-0.0189 (5.88)	0.0634 (2.29)	6%	-0.0131 (4.87)	0.0708 (-0.54)	-0.0121 (2.73)	0.0292 (-1.68)	9%	-0.0010 (4.71)	0.0827 (-0.04)	-0.0233 (3.11)	0.0881 (-1.61)	-0.0175 (2.58)	0.0325 (-2.24)	(2.85)
1999	251	3%	0.0190 (6.81)	-0.0474 (2.71)	5%	0.0013 (4.71)	-0.0231 (0.11)	0.0114 (-2.13)	-0.0677 (-3.40)	7%	0.0071 (3.77)	-0.0261 (0.47)	0.0195 (-2.38)	-0.0785 (2.68)	-0.0554 (-3.80)	-0.0215 (-1.30)	(-1.69)
2000	251	1%	-0.0335 (2.63)	0.0114 (-1.63)	3%	-0.0389 (2.20)	0.0130 (-1.77)	-0.0040 (1.83)	0.1019 (-1.45)	4%	-0.0323 (1.94)	0.0228 (-1.40)	0.0127 (1.17)	-0.0662 (1.75)	-0.0049 (-1.36)	0.1247 (-1.73)	(1.71)
2001	247	2%	0.0134 (4.31)	-0.0421 (1.99)	5%	-0.0002 (4.05)	-0.0791 (-0.02)	0.0308 (-2.91)	0.0826 (1.78)	6%	0.0066 (3.19)	-0.0775 (0.48)	0.0373 (-2.86)	0.0198 (2.13)	-0.0963 (1.71)	0.0486 (-1.35)	(1.31)
2002	251	2%	0.0193 (4.84)	-0.0433 (2.09)	5%	-0.0017 (3.97)	0.0115 (-0.08)	-0.0096 (1.91)	-0.0450 (-2.31)	7%	0.0430 (3.69)	0.0127 (1.17)	0.2062 (2.12)	-0.0152 (1.81)	-0.0098 (-1.93)	-0.0457 (-2.11)	(-2.37)
EntirePeriod	3026	0%	0.0050 (7.01)	-0.0108 (2.86)	1%	0.0032 (5.58)	0.0047 (1.73)	0.0243 (2.00)	-0.0111 (-2.69)	1%	-0.0111 (5.14)	0.0009 (-2.15)	0.0059 (2.19)	0.0207 (2.47)	-0.0078 (-1.83)	0.0077 (2.31)	(2.31)

Table 6.4 Daily Stepwise Linear Regressions: VIXFirstDiff = Categories, No Lag

VIXFirstDiff

Period	Total Size	1-Variable			3-Variable					5-Variable						
		R-Square	Intercept		R-Square	Intercept					R-Square	Intercept				
1991	253	3%	0.3942 (7.46)	Milit -1.3395 (-2.73)	7%	-9.0774 (6.46)	Active 0.8619 (2.20)	Increase 3.6917 (2.48)	Milit -1.3990 (-2.88)	10%	-6.9132 (5.68)	0.7165 (1.93)	3.3320 (2.17)	Legal 1.6072 (2.18)	Milit -1.4455 (-2.78)	RcEthic -6.4215 (-2.89)
1992	255	1%	0.4869 (3.09)	Yes -1.83 (-1.76)	3%	0.3014 (2.64)	Milit -1.3877 (-1.39)	Pleasur -4.1797 (-1.65)	Yes 6.5342 (1.80)	4%	2.6149 (2.31)	-1.4137 (-1.42)	Need -2.8470 (-1.47)	Pleasur -3.3114 (-1.29)	Unrst -0.8174 (-1.25)	Yes 7.3425 (2.02)
1993	252	1%	2.0452 (1.81)	Positiv -0.4756 (-1.82)	4%	3.8382 (3.44)	Comple 2.9760 (1.89)	Passive -0.7007 (-2.36)	Positiv -0.6318 (-2.36)	5%	2.7714 (2.47)	0.2285 (1.29)	2.8390 (1.06)	-0.7545 (-1.00)	-0.6964 (-1.74)	Passive -0.6947 (-2.45)
1994	251	4%	2.0227 (11.49)	Increase -3.8454 (-3.39)	9%	-0.4645 (8.52)	-3.8602 (-3.44)	Need -7.9858 (-3.00)	Positiv 1.0576 (2.52)	13%	0.9137 (7.19)	5.5101 (0.45)	-4.1280 (-3.72)	-9.0942 (-3.34)	-9.2207 (-3.00)	Positiv 0.8326 (1.96)
1995	251	3%	2.9178 (8.46)	Unrst -1.3655 (-2.91)	7%	3.8656 (6.11)	-0.5125 (3.29)	Need 2.2938 (1.84)	Unrst -1.3520 (-2.87)	9%	6.2921 (5.13)	-2.8313 (4.17)	-0.4969 (-1.80)	2.4677 (-2.13)	-0.6328 (-2.05)	Overst -1.2894 (-2.95)
1996	252	2%	3.8716 (4.12)	Passive 1.3622 (2.03)	4%	-3.2543 (3.23)	1.6324 (-1.61)	1.7146 (1.63)	Weak -1.5661 (-1.71)	6%	-1.2128 (3.12)	1.8648 (-0.51)	1.7321 (1.86)	SureLw -2.4077 (2.42)	Weak -1.4192 (-2.09)	Yes 9.2848 (1.56)
1997	252	2%	-0.8176 (5.96)	Feel 39.5558 (2.44)	6%	1.0499 (5.35)	43.6092 (0.62)	-4.4380 (-2.42)	10.0022 (2.32)	10%	-0.2601 (5.39)	47.8693 (-0.12)	-3.8124 (3.01)	5.1060 (-2.09)	12.9798 (-2.09)	Pain -12.2217 (-2.73)
1998	251	3%	6.7036 (6.97)	Virtue -4.2790 (-2.64)	7%	5.7879 (5.82)	7.0210 (1.89)	-13.2222 (-2.35)	-6.0549 (-3.55)	10%	18.0119 (5.18)	-10.0691 (3.23)	6.3652 (2.40)	-2.6256 (-2.00)	-17.9426 (-3.09)	Virtue -4.9216 (-2.81)
1999	250	1%	1.7537 (8.70)	WlbPsc -14.0474 (-2.95)	6%	4.2436 (5.65)	-1.7616 (2.35)	5.4845 (-2.66)	-12.6766 (-2.66)	9%	3.0955 (4.59)	6.5512 (1.41)	-1.4779 (1.81)	-1.0108 (-1.67)	5.8634 (2.03)	WlbPsc -14.2819 (-2.95)
2000	251	2%	3.7576 (4.38)	SureLw -4.1303 (-2.09)	4%	2.7389 (3.00)	-0.5693 (0.58)	0.4819 (-2.04)	-4.7610 (-1.98)	5%	-5.6090 (2.41)	-0.5951 (1.08)	-17.9563 (-1.33)	-0.0129 (-1.08)	6.2157 (1.62)	SureLw -3.8190 (-1.54)
2001	247	2%	-4.6641 (5.40)	Virtue 2.8875 (2.32)	4%	-12.1445 (3.71)	1.0973 (-2.66)	-4.6607 (-1.31)	2.3034 (1.83)	6%	-15.4518 (2.91)	1.1481 (3.16)	-5.0002 (2.30)	3.1539 (-1.34)	4.3053 (1.55)	Virtue 2.5374 (1.98)
2002	251	1%	1.2269 (5.48)	No -31.4586 (-2.34)	5%	-8.1006 (4.05)	-27.7929 (-2.07)	5.1108 (2.11)	0.6382 (2.06)	7%	-4.2928 (3.55)	-2.3130 (-0.90)	-4.8708 (-2.10)	-27.4922 (-2.06)	5.2157 (2.50)	Strong 0.4971 (1.34)
EntirePeriod	3027	0%	1.2883 (10.06)	SureLw -1.4126 (-3.17)	1%	-0.3847 (7.18)	-0.8861 (-0.45)	0.1708 (-2.66)	-1.3838 (-3.10)	1%	0.0579 (5.89)	-0.1703 (0.06)	-1.1067 (-2.44)	-1.7752 (-3.16)	0.2939 (3.43)	SureLw -1.5693 (-3.48)

Table 6.5 Daily Stepwise Linear Regressions: BondReturn = Categories, No Lag

BondReturn

Period	Total Size	1-Variable			3-Variable					5-Variable						
		R-Square	Intercept		R-Square	Intercept					R-Square	Intercept				
1991	253	2%	0.0024 (6.11)	No -0.0759 (-2.47)	4%	-0.0106 (3.88)	-0.0201 (-1.15)	-0.0783 (-1.70)	0.0055 (1.81)	6%	-0.0032 (3.26)	-0.0203 (-0.31)	-0.0836 (-1.71)	0.0057 (-2.73)	0.0227 (1.83)	Virtue -0.0067 (-1.79)
1992	255	3%	0.0204 (8.80)	Negativ -0.0071 (-2.97)	7%	0.0206 (6.61)	0.0103 (2.85)	-0.0061 (2.60)	-0.0382 (-3.23)	9%	0.0240 (5.04)	0.0126 (1.84)	-0.0065 (3.08)	-0.0426 (-2.69)	0.0125 (3.37)	WlbPsc -0.0093 (-2.05)
1993	253	1%	0.0048 (2.94)	RcEthic -0.0107 (-1.72)	4%	0.0059 (3.09)	0.0221 (1.57)	-0.0133 (-2.11)	-0.0345 (-2.38)	5%	-0.0039 (2.56)	0.0023 (-0.52)	0.0225 (1.27)	-0.0129 (1.68)	0.0326 (-2.05)	Yes 0.0305 (1.62)
1994	253	2%	-0.0053 (5.50)	No 0.1707 (2.35)	4%	-0.0017 (3.81)	0.0135 (-0.29)	0.1713 (1.73)	-0.0230 (-1.94)	7%	-0.0006 (3.49)	0.0133 (-0.08)	0.0350 (1.67)	0.1599 (1.50)	-0.0222 (2.21)	WlbPsc -0.0617 (-2.11)
1995	252	1%	0.0057 (2.65)	Try -0.0356 (-1.53)	3%	0.0048 (2.76)	-0.0166 (0.74)	0.0390 (1.84)	-0.0473 (-2.11)	5%	-0.0150 (2.40)	0.0951 (-0.95)	-0.0159 (1.41)	0.0466 (-1.55)	0.0640 (2.17)	Try -0.0519 (-2.28)
1996	254	3%	-0.0049 (9.00)	Feel 0.2300 (3.00)	8%	0.0144 (7.04)	0.2540 (2.43)	-0.0694 (3.36)	-0.0457 (-1.97)	10%	0.0187 (5.76)	-0.0292 (2.22)	0.2581 (-1.65)	0.0453 (3.44)	-0.0748 (-3.05)	WlbPsc -0.0657 (-2.56)
1997	253	3%	0.0217 (6.74)	Negativ -0.0061 (-2.60)	6%	0.0287 (5.13)	0.0256 (3.14)	-0.0082 (-3.38)	-0.0367 (-2.16)	9%	0.0470 (4.83)	0.0234 (3.54)	0.0265 (2.12)	-0.0094 (-2.58)	-0.0018 (-1.97)	Strong -0.0203 (-2.03)
1998	253	2%	0.0236 (4.25)	Positiv -0.0057 (-2.06)	3%	0.0249 (2.35)	-0.0190 (-2.16)	-0.0127 (-1.26)	-0.0042 (-1.44)	3%	0.0203 (1.75)	-0.0203 (-1.53)	-0.0130 (-1.16)	0.0072 (-1.24)	-0.0181 (0.93)	Positiv -0.0305 (-1.17)
1999	253	1%	-0.0057 (2.13)	Need 0.0219 (1.48)	2%	-0.0007 (1.90)	0.0235 (-0.06)	0.0034 (1.34)	-0.0041 (-1.45)	4%	0.0137 (2.08)	-0.0131 (-0.95)	0.0269 (-1.51)	0.0060 (2.14)	-0.0048 (-1.57)	RcEthic -0.0224 (-1.82)
2000	253	2%	-0.0034 (4.04)	Politi@ 0.0021 (2.01)	4%	0.0109 (3.25)	0.0181 (1.38)	-0.0059 (-2.09)	0.0018 (1.68)	8%	0.0019 (3.03)	0.0205 (0.21)	-0.0060 (-1.96)	0.0012 (1.07)	0.0125 (1.49)	Try 0.0350 (1.87)
2001	253	2%	0.0199 (4.79)	Positiv -0.0049 (-2.18)	4%	0.0249 (3.67)	0.0240 (2.37)	-0.0257 (-2.84)	-0.0065 (-2.84)	6%	-0.0012 (3.16)	0.0034 (-0.07)	0.0228 (1.93)	0.0050 (-1.91)	-0.0306 (-1.65)	Positiv -0.0080 (-3.29)
2002	253	3%	0.0355 (6.81)	Active -0.0037 (-2.62)	5%	0.0640 (4.27)	-0.0041 (3.38)	-0.0048 (-2.78)	-0.0099 (-1.32)	6%	0.0636 (3.33)	-0.0047 (-3.31)	-0.0141 (-2.97)	0.0029 (-1.83)	-0.0039 (-1.09)	SureLw -0.0102 (-1.28)
EntirePeriod	3049	0%	0.0016 (4.40)	Goal -0.0048 (-2.10)	0%	0.0015 (3.13)	0.0026 (1.24)	-0.0047 (-1.99)	-0.0091 (-1.97)	0%	0.0038 (2.81)	0.0027 (1.70)	0.0266 (1.73)	-0.0054 (-2.27)	-0.0029 (-1.50)	WlbPsc -0.0097 (-2.10)

Table 6.6 Daily Stepwise Linear Regressions: DEMEuroUSDReturn = Categories, No Lag

DemEuroUSDReturn																
Period	Total Size	1-Variable			3-Variable					5-Variable						
1991	253	R-Square	Intercept	Econ@	R-Square	Intercept	Econ@	Legal	Vice	R-Square	Intercept	Decreas	Econ@	Legal	Try	Vice
		1%	-0.0153	0.0022	3%	-0.0299	0.0025	0.0050	0.0074	6%	-0.0209	-0.0231	0.0030	0.0039	-0.0401	0.0103
1992	255	R-Square	Intercept	Positiv	R-Square	Intercept	Goal	Positiv	Try	R-Square	Intercept	Exprsv	Goal	No	Positiv	Try
		3%	0.0389	-0.0090	5%	0.0295	0.0148	-0.0094	0.0362	7%	0.0223	0.0099	0.0166	0.0691	-0.0096	0.0329
1993	253	R-Square	Intercept	Pain	R-Square	Intercept	Compleat	Pain	Virtue	R-Square	Intercept	Compleat	Feel	Goal	Pain	Virtue
		3%	0.0078	-0.0403	5%	0.0047	-0.0198	-0.0385	0.0057	7%	-0.0012	-0.0203	0.0807	0.0094	-0.0378	0.0063
1994	253	R-Square	Intercept	AffTot	R-Square	Intercept	Active	AffTot	Negativ	R-Square	Intercept	Active	AffTot	Negativ	RcEthic	Weak
		2%	0.0037	-0.0089	4%	-0.0081	0.0027	-0.0104	-0.0041	5%	-0.0130	0.0027	-0.0110	-0.0065	0.0107	0.0048
1995	252	R-Square	Intercept	RcEthic	R-Square	Intercept	No	RcEthic	Undrst	R-Square	Intercept	Feel	Milit	No	RcEthic	Undrst
		1%	0.0066	-0.0160	3%	-0.0100	-0.0721	-0.0169	0.0091	5%	-0.0166	0.0886	0.0118	-0.0777	-0.0168	0.0103
1996	254	R-Square	Intercept	Compleat	R-Square	Intercept	Compleat	Increas	Ovrst	R-Square	Intercept	Compleat	Increas	Ovrst	Politi@	Weak
		2%	-0.0057	0.0186	6%	0.0033	0.0188	0.0115	-0.0042	9%	0.0143	0.0192	0.0113	-0.0043	-0.0017	-0.0049
1997	253	R-Square	Intercept	RspTot	R-Square	Intercept	Pain	RspTot	Virtue	R-Square	Intercept	No	Pain	RspTot	Virtue	Weak
		3%	0.0086	-0.0191	5%	0.0212	-0.0287	-0.0190	-0.0046	6%	0.0162	0.0321	-0.0338	-0.0202	-0.0056	0.0044
1998	253	R-Square	Intercept	If	R-Square	Intercept	If	Legal	Politi@	R-Square	Intercept	AffTot	If	Legal	Politi@	Positiv
		3%	0.0123	-0.0140	7%	0.0075	-0.0113	0.0082	-0.0042	10%	0.0186	0.0107	-0.0118	0.0100	-0.0041	-0.0043
1999	253	R-Square	Intercept	Politi@	R-Square	Intercept	Need	Politi@	RspTot	R-Square	Intercept	Compleat	Fail	Need	Politi@	RspTot
		2%	0.0073	-0.0042	5%	0.0077	-0.0230	-0.0041	0.0123	8%	0.0027	0.0210	-0.0235	-0.0210	-0.0041	0.0148
2000	253	R-Square	Intercept	Try	R-Square	Intercept	Means	No	Try	R-Square	Intercept	Means	No	Think	Try	Virtue
		2%	0.0097	-0.0576	4%	0.0018	0.0036	-0.0727	-0.0545	5%	-0.0048	0.0034	-0.0834	-0.0701	-0.0580	0.0060
2001	253	R-Square	Intercept	Try	R-Square	Intercept	Increas	Pleasur	Try	R-Square	Intercept	Increas	Need	Pleasur	Politi@	Try
		2%	-0.0070	0.0454	6%	-0.0228	0.0192	0.0354	0.0584	8%	-0.0370	0.0197	0.0321	0.0373	0.0043	0.0480
2002	253	R-Square	Intercept	Yes	R-Square	Intercept	Feel	Try	Yes	R-Square	Intercept	Feel	Politi@	RspTot	Try	Yes
		2%	0.0040	-0.0610	6%	-0.0022	0.1059	0.0281	-0.0796	7%	0.0052	0.1113	-0.0023	-0.0084	0.0308	-0.0808
EntirePeriod	3049	R-Square	Intercept	Pain	R-Square	Intercept	Fail	Goal	Pain	R-Square	Intercept	Fail	Goal	Pain	Pleasur	Politi@
		1%	0.0027	-0.0133	1%	0.0013	-0.0113	0.0076	-0.0098	1%	0.0015	-0.0104	0.0082	-0.0107	0.0062	-0.0007

Table 6.7 Daily Stepwise Linear Regressions: YenUSDReturn = Categories, No Lag

YenUSDReturn																
Period	Total Size	1-Variable			3-Variable					5-Variable						
1991	253	R-Square	Intercept	No	R-Square	Intercept	Goal	No	RspTot	R-Square	Intercept	Active	Exprsv	Goal	No	RspTot
		3%	-0.0030	0.0847	6%	-0.0118	0.0115	0.0946	0.0123	8%	-0.0309	0.0024	-0.0063	0.0109	0.0817	0.0157
1992	255	R-Square	Intercept	Passive	R-Square	Intercept	Goal	Passive	Vice	R-Square	Intercept	Decreas	Goal	Passive	Vice	Weak
		1%	0.0133	-0.0043	3%	0.0071	0.0082	-0.0048	0.0062	4%	0.0124	0.0100	0.0074	-0.0041	0.0070	-0.0061
1993	253	R-Square	Intercept	Think	R-Square	Intercept	Negativ	Pleasur	Think	R-Square	Intercept	AffTot	Goal	Negativ	Pleasur	Think
		2%	0.0021	-0.1255	5%	0.0176	-0.0043	-0.0255	-0.1193	7%	0.0104	0.0087	0.0107	-0.0042	-0.0321	-0.1101
1994	253	R-Square	Intercept	AffTot	R-Square	Intercept	AffTot	Compleat	RcEthic	R-Square	Intercept	AffTot	Arousal	Compleat	Need	RcEthic
		2%	0.0039	-0.0092	4%	0.0026	-0.0085	-0.0177	0.0152	6%	0.0049	-0.0097	0.0198	-0.0185	-0.0210	0.0152
1995	252	R-Square	Intercept	RcEthic	R-Square	Intercept	RcEthic	Undrst	Vice	R-Square	Intercept	Milit	RcEthic	Undrst	Vice	Weak
		3%	0.0116	-0.0267	6%	-0.0172	-0.0225	0.0155	-0.0090	8%	-0.0259	0.0102	-0.0229	0.0134	-0.0120	0.0085
1996	254	R-Square	Intercept	Politi@	R-Square	Intercept	Ovrst	Pain	Politi@	R-Square	Intercept	Means	Ovrst	Pain	Politi@	Strong
		2%	0.0046	-0.0024	5%	0.0251	-0.0045	-0.0198	-0.0028	8%	0.0219	-0.0047	-0.0042	-0.0207	-0.0036	0.0016
1997	253	R-Square	Intercept	Weak	R-Square	Intercept	Compleat	SureLw	Weak	R-Square	Intercept	Compleat	Pain	Pleasur	SureLw	Weak
		2%	-0.0130	0.0089	3%	0.0035	-0.0179	-0.0146	0.0105	5%	0.0029	-0.0206	-0.0222	0.0330	-0.0154	0.0119
1998	253	R-Square	Intercept	Weak	R-Square	Intercept	Increas	Legal	Weak	R-Square	Intercept	Decreas	Feel	Increas	Legal	Weak
		4%	0.0281	-0.0189	7%	0.0242	-0.0175	0.0102	-0.0186	9%	0.0259	-0.0356	-0.2109	-0.0172	0.0099	-0.0123
1999	253	R-Square	Intercept	Try	R-Square	Intercept	No	Positiv	Try	R-Square	Intercept	If	No	Positiv	SureLw	Try
		2%	-0.0091	0.0530	5%	0.0124	0.0817	-0.0066	0.0737	8%	0.0181	-0.0173	0.0727	-0.0081	0.0171	0.0797
2000	253	R-Square	Intercept	Vice	R-Square	Intercept	Goal	Try	Vice	R-Square	Intercept	Exprsv	Goal	Try	Vice	Virtue
		2%	0.0060	-0.0090	4%	0.0083	0.0099	-0.0324	-0.0092	6%	0.0013	-0.0035	0.0096	-0.0352	-0.0079	0.0057
2001	253	R-Square	Intercept	Arousal	R-Square	Intercept	Arousal	Fail	Vice	R-Square	Intercept	Arousal	Fail	Feel	Ovrst	Vice
		1%	0.0038	-0.0190	4%	0.0061	-0.0355	-0.0246	0.0062	6%	0.0191	-0.0370	-0.0270	0.1056	-0.0046	0.0082
2002	253	R-Square	Intercept	Yes	R-Square	Intercept	Exprsv	Positiv	Yes	R-Square	Intercept	Exprsv	Fail	Positiv	Undrst	Yes
		3%	0.0053	-0.0731	6%	-0.0077	-0.0046	0.0036	-0.0695	8%	0.0061	-0.0049	-0.0199	0.0037	-0.0056	-0.0697
EntirePeriod	3049	R-Square	Intercept	Goal	R-Square	Intercept	Exprsv	Fail	Goal	R-Square	Intercept	AffTot	Arousal	Exprsv	Fail	Goal
		0%	-0.0018	0.0052	0%	0.0004	-0.0014	-0.0103	0.0059	1%	0.0000	0.0034	-0.0066	-0.0018	-0.0106	0.0069

Table 6.8 Daily Stepwise Linear Regressions: OilReturn = Categories, No Lag

OilReturn																
Period	Total Size	1-Variable			3-Variable					5-Variable						
1991	251	R-Square	Intercept	SureLw	R-Square	Intercept	Comple	Pain	SureLw	R-Square	Intercept	AffTot	Comple	Means	Pain	SureLw
		3%	0.1216	-0.1339	6%	0.2131	-0.1883	-0.1850	-0.1254	9%	0.3847	-0.0844	-0.1990	-0.0365	-0.1891	-0.1482
1992	252	R-Square	Intercept	Passive	R-Square	Intercept	AffTot	Means	Passive	R-Square	Intercept	AffTot	Econ@	Means	Passive	Virtue
		2%	-0.0538	0.0180	5%	-0.0874	-0.0323	0.0100	0.0238	8%	-0.0959	-0.0379	-0.0052	0.0170	0.0217	0.0203
1993	250	R-Square	Intercept	Feel	R-Square	Intercept	Feel	Weak	WibPsyc	R-Square	Intercept	Feel	Need	Undrst	Weak	WibPsyc
		6%	0.0119	-0.6142	9%	-0.0287	-0.6243	0.0184	0.0935	11%	0.0078	-0.5944	0.0592	-0.0285	0.0249	0.0836
1994	250	R-Square	Intercept	RspTot	R-Square	Intercept	Pain	RspTot	WibPsyc	R-Square	Intercept	Econ@	If	Pain	RspTot	WibPsyc
		2%	-0.0254	0.0622	5%	-0.0240	-0.0942	0.0604	0.1439	7%	-0.0343	0.0056	-0.0372	-0.0843	0.0594	0.1889
1995	250	R-Square	Intercept	Fail	R-Square	Intercept	Active	Fail	Persist	R-Square	Intercept	Active	Exprsv	Fail	Persist	Positiv
		3%	0.0176	-0.1176	6%	-0.0692	0.0075	-0.1208	0.0970	8%	-0.0570	0.0093	0.0225	-0.1049	0.0987	-0.0106
1996	251	R-Square	Intercept	AffTot	R-Square	Intercept	AffTot	Ovrst	Polit@	R-Square	Intercept	AffTot	Econ@	Means	Ovrst	Polit@
		4%	0.0370	-0.0774	9%	0.2225	-0.0908	-0.0413	-0.0182	12%	0.2209	-0.0812	0.0128	-0.0383	-0.0360	-0.0128
1997	251	R-Square	Intercept	Strong	R-Square	Intercept	AffTot	Feel	Strong	R-Square	Intercept	AffTot	Feel	Goal	Passive	Strong
		1%	-0.0683	0.0062	3%	-0.1036	0.0263	-0.2308	0.0089	4%	-0.0522	0.0264	-0.2372	0.0313	-0.0127	0.0065
1998	250	R-Square	Intercept	Milit	R-Square	Intercept	Goal	If	Milit	R-Square	Intercept	Fail	Goal	If	Milit	Need
		2%	0.0161	-0.0912	5%	0.1206	-0.1188	-0.0733	-0.1023	7%	0.0857	0.1494	-0.1193	-0.1044	-0.0980	0.1644
1999	249	R-Square	Intercept	Milit	R-Square	Intercept	Milit	Vice	Weak	R-Square	Intercept	Legal	Milit	Ovrst	Vice	Weak
		2%	-0.0082	0.0505	7%	0.0458	0.0409	0.0459	0.0557	9%	-0.0563	0.0209	0.0463	0.0276	0.0364	-0.0648
2000	249	R-Square	Intercept	Decreas	R-Square	Intercept	Arousal	Decreas	Positiv	R-Square	Intercept	Active	Arousal	Decreas	Polit@	Positiv
		2%	0.0234	-0.1288	5%	-0.0798	-0.1107	-0.0952	0.0283	7%	0.0458	-0.0167	-0.1571	-0.1095	0.0093	0.0330
2001	250	R-Square	Intercept	Active	R-Square	Intercept	Active	EMOT	Try	R-Square	Intercept	Active	Arousal	Econ@	EMOT	Try
		1%	0.1127	-0.0127	3%	0.1386	-0.0163	-0.0679	0.1770	5%	0.1437	-0.0118	0.1001	-0.0062	-0.1311	0.1357
2002	250	R-Square	Intercept	Weak	R-Square	Intercept	RcEthic	Undrst	Weak	R-Square	Intercept	Negativ	Undrst	RcEthic	Undrst	Weak
		3%	0.0697	-0.0378	7%	0.0081	-0.0384	0.0380	-0.0375	9%	-0.0616	0.0166	0.0157	-0.0566	0.0317	-0.0526
EntirePeriod	3014	R-Square	Intercept	Exprsv	R-Square	Intercept	Exprsv	Pain	Virtue	R-Square	Intercept	Exprsv	If	Means	Pain	Virtue
		0%	-0.0055	0.0094	0%	-0.0123	0.0082	-0.0331	0.0089	1%	0.0122	0.0046	-0.0134	-0.0043	-0.0317	0.0104
		(6.03)	(-2.28)	(2.45)	(4.58)	(-1.71)	(2.15)	(-2.22)	(2.04)	(3.77)	(0.84)	(1.09)	(-1.80)	(-1.58)	(-1.97)	(2.15)

Table 6.9 Daily Stepwise Linear Regressions: GoldReturn = Categories, No Lag

GoldReturn																
Period	Total Size	1-Variable			3-Variable					5-Variable						
1991	252	R-Square	Intercept	Legal	R-Square	Intercept	Exprsv	Increas	Legal	R-Square	Intercept	Exprsv	Increas	Legal	Need	Polit@
		2%	-0.0109	0.0077	4%	-0.0165	-0.0107	0.0184	0.0095	6%	-0.0144	-0.0121	0.0176	0.0113	0.0336	-0.0061
1992	253	R-Square	Intercept	WibPsyc	R-Square	Intercept	Need	RcEthic	WibPsyc	R-Square	Intercept	Comple	Need	RcEthic	Strong	WibPsyc
		2%	0.0047	-0.0419	7%	-0.0119	0.0377	0.0167	-0.0432	9%	0.0107	0.0222	0.0389	0.0198	-0.0028	-0.0468
1993	251	R-Square	Intercept	If	R-Square	Intercept	EMOT	If	Yes	R-Square	Intercept	EMOT	If	No	SureLw	Yes
		3%	0.0255	-0.0273	6%	0.0109	0.0337	-0.0291	0.0698	8%	0.0258	0.0356	-0.0209	-0.1270	-0.0219	0.0896
1994	252	R-Square	Intercept	Pleasur	R-Square	Intercept	Decreas	Passive	Pleasur	R-Square	Intercept	Decreas	Econ@	Passive	Pleasur	Positiv
		4%	-0.0084	0.0657	8%	0.0128	0.0306	-0.0098	0.0754	12%	0.0182	0.0415	-0.0035	-0.0124	0.0563	0.0060
1995	251	R-Square	Intercept	Vice	R-Square	Intercept	Increas	Strong	Vice	R-Square	Intercept	Increas	Negativ	No	Strong	Vice
		3%	-0.0052	0.0073	5%	-0.0166	-0.0116	0.0017	0.0051	8%	-0.0084	-0.0137	-0.0044	-0.0547	0.0019	0.0123
1996	253	R-Square	Intercept	Think	R-Square	Intercept	If	Pleasur	Think	R-Square	Intercept	If	Negativ	Passive	Pleasur	Think
		2%	0.0011	-0.0557	3%	-0.0071	0.0066	0.0191	-0.0573	5%	-0.0016	0.0087	0.0024	-0.0051	0.0228	-0.0485
1997	253	R-Square	Intercept	If	R-Square	Intercept	If	Strong	WibPsyc	R-Square	Intercept	If	Strong	Undrst	WibPsyc	
		2%	0.0147	-0.0179	4%	-0.0206	-0.0173	0.0027	0.0476	6%	-0.0049	-0.0144	-0.0164	0.0031	-0.0099	0.0587
1998	253	R-Square	Intercept	Milit	R-Square	Intercept	Milit	Passive	Positiv	R-Square	Intercept	EMOT	Milit	Passive	Positiv	SureLw
		2%	0.0043	-0.0232	5%	0.0588	-0.0263	-0.0077	-0.0076	7%	0.0576	0.0242	-0.0309	-0.0099	-0.0095	0.0093
1999	253	R-Square	Intercept	Polit@	R-Square	Intercept	Econ@	Pleasur	Polit@	R-Square	Intercept	Comple	Econ@	Passive	Pleasur	Polit@
		1%	-0.0114	0.0069	5%	-0.0490	0.0032	0.0982	0.0092	6%	-0.0259	0.0430	0.0022	-0.0098	0.0883	0.0087
2000	253	R-Square	Intercept	Try	R-Square	Intercept	EMOT	Pain	Try	R-Square	Intercept	EMOT	Exprsv	Pain	Pleasur	Try
		1%	-0.0094	0.0573	3%	-0.0145	0.0396	-0.0381	0.0556	5%	-0.0131	0.0522	0.0073	-0.0443	-0.0588	0.0508
2001	251	R-Square	Intercept	Virtue	R-Square	Intercept	Exprsv	Persist	Virtue	R-Square	Intercept	Exprsv	Need	Persist	Strong	Virtue
		4%	-0.0305	0.0190	8%	-0.0275	0.0094	-0.0487	0.0196	10%	-0.0528	0.0140	-0.0332	-0.0531	0.0026	0.0220
2002	253	R-Square	Intercept	Vice	R-Square	Intercept	Ovrst	Polit@	Vice	R-Square	Intercept	Arousal	Ovrst	Polit@	Vice	WibPsyc
		2%	0.0161	-0.0179	5%	-0.0174	0.0070	0.0054	-0.0203	7%	-0.0178	0.0430	0.0073	0.0055	-0.0229	-0.0476
EntirePeriod	3039	R-Square	Intercept	Virtue	R-Square	Intercept	Exprsv	Polit@	Virtue	R-Square	Intercept	EMOT	Exprsv	Milit	Polit@	Virtue
		0%	-0.0057	0.0036	1%	-0.0105	0.0035	0.0016	0.0034	1%	-0.0113	0.0083	0.0030	-0.0038	0.0015	0.0030
		(7.04)	(-2.67)	(2.65)	(6.04)	(-4.00)	(2.76)	(2.57)	(2.54)	(5.14)	(-4.20)	(2.13)	(2.38)	(-2.15)	(2.43)	(2.17)

Table 6.10 Daily Stepwise Linear Regressions: CBOEPutCall = Categories, No Lag

CBOEPutCall																
Period	Total Size	1-Variable			3-Variable					5-Variable						
		R-Square	Intercept	Goal	R-Square	Intercept	SureLw	Virtue	R-Square	Intercept	Active	If	SureLw	Virtue		
1991	252	3%	0.9083	-0.5074	9%	0.9146	-0.5093	-0.5122	0.2982	14%	1.6316	-0.1170	-0.4084	0.4229	-0.7131	0.3486
		(8.74)	(15.56)	(-2.96)	(7.77)	(3.89)	(-3.00)	(-2.77)	(2.87)	(7.98)	(3.89)	(-2.76)	(-2.43)	(2.92)	(-3.61)	(3.42)
1992	253	2%	1.1144	-0.3844	5%	0.8985	0.6858	0.5776	-0.3925	7%	0.9667	0.7018	0.7270	-0.5549	-0.3215	-2.1955
		(4.55)	(6.73)	(-2.13)	(3.93)	(4.91)	(1.96)	(1.66)	(-2.20)	(3.53)	(5.11)	(2.02)	(2.07)	(-1.53)	(-1.79)	(-1.86)
1993	252	2%	0.6481	0.4685	5%	1.0287	0.4890	-0.2432	-0.7788	7%	1.0958	-0.2255	0.4330	-0.2437	-0.8848	0.5841
		(5.22)	(13.34)	(2.28)	(4.63)	(7.01)	(2.40)	(-1.83)	(-2.15)	(3.90)	(6.69)	(-2.02)	(2.07)	(-1.85)	(-2.43)	(1.53)
1994	251	2%	1.1162	-0.5879	6%	1.5526	-0.7173	-0.4753	-0.7061	8%	1.4990	-1.1207	0.8845	-0.4832	-0.6895	-0.2296
		(5.70)	(10.38)	(-2.39)	(4.81)	(8.49)	(-1.58)	(-2.38)	(-2.84)	(4.48)	(7.07)	(-2.32)	(2.43)	(-2.41)	(-2.77)	(-1.76)
1995	251	4%	0.3657	0.1525	8%	0.5689	-0.2631	0.1440	-1.7370	9%	0.6018	-0.3130	0.5257	0.1374	-1.8913	-0.5546
		(11.63)	(2.92)	(3.41)	(6.76)	(3.99)	(-2.15)	(3.25)	(-1.94)	(5.09)	(3.54)	(-2.50)	(1.67)	(3.10)	(-2.10)	(-1.35)
1996	253	1%	0.9782	-0.5900	7%	0.6966	0.7753	-0.5107	0.8431	8%	0.4729	0.7271	0.3580	-0.3748	-0.9687	0.9877
		(8.82)	(11.66)	(-2.97)	(6.62)	(5.81)	(2.87)	(-2.59)	(2.19)	(6.00)	(2.46)	(2.71)	(2.29)	(-1.88)	(-2.31)	(2.53)
1997	252	6%	0.1333	0.1948	12%	0.2989	0.1572	0.4828	-0.3266	15%	0.5946	-0.8091	0.1794	-0.3723	0.4652	-0.3188
		(16.45)	(1.01)	(4.06)	(11.17)	(1.34)	(3.30)	(2.74)	(-3.57)	(8.68)	(2.48)	(-2.22)	(3.76)	(-2.01)	(2.87)	(-3.52)
1998	251	13%	-0.1522	0.5397	22%	0.4620	-1.3471	-3.0368	0.4585	25%	0.4328	-1.2312	0.9337	-3.2198	0.3532	-0.3391
		(38.78)	(-1.14)	(6.23)	(23.34)	(2.66)	(-4.89)	(-2.45)	(5.44)	(16.68)	(2.52)	(-4.49)	(2.51)	(-6.22)	(3.77)	(-2.22)
1999	251	3%	0.2694	0.1129	7%	0.7545	-0.4654	0.0946	-0.1822	10%	0.9232	-0.4555	-0.3274	0.0624	-0.1758	-0.1758
		(8.93)	(2.63)	(2.95)	(6.84)	(4.19)	(-2.08)	(2.51)	(-2.66)	(5.25)	(6.85)	(-2.55)	(1.56)	(-2.32)	(1.55)	(-2.54)
2000	251	17%	0.2906	1.5409	33%	1.8615	1.4868	-0.0801	-0.2568	38%	1.2603	1.1115	-0.0631	0.5006	0.1913	-0.2931
		(50.94)	(7.47)	(7.14)	(40.24)	(7.92)	(7.17)	(-5.78)	(-5.18)	(30.38)	(4.73)	(5.23)	(-4.53)	(2.48)	(3.37)	(-6.05)
2001	246	9%	0.4707	1.1035	12%	0.4146	-0.4354	0.6642	0.2169	15%	0.2003	-0.5343	-0.4820	0.1234	0.5164	0.2731
		(23.27)	(9.80)	(4.82)	(11.39)	(2.42)	(-2.56)	(2.50)	(2.09)	(8.46)	(1.00)	(-2.44)	(-2.22)	(1.79)	(1.85)	(2.48)
2002	251	3%	0.6167	0.8239	7%	0.6119	-0.5194	0.7211	0.8091	8%	0.8276	-0.1434	-0.4510	-0.5999	0.8121	0.8481
		(6.56)	(8.17)	(2.56)	(6.05)	(5.71)	(-2.27)	(2.65)	(2.56)	(4.48)	(5.45)	(-1.41)	(-1.61)	(-2.52)	(2.86)	(2.87)
EntirePeriod	3025	10%	0.0775	0.4065	14%	-0.3772	0.1037	0.1412	0.3888	15%	-0.3914	0.0680	0.0940	0.1649	0.2793	-0.5307
		(338.66)	(2.21)	(18.40)	(161.65)	(-5.95)	(8.73)	(5.59)	(17.83)	(104.75)	(-5.84)	(4.24)	(7.79)	(6.37)	(8.67)	(-3.91)

Table 6.11 Daily Stepwise Linear Regressions: NYSEAdvanceDecline = Categories, No Lag

NYSEAdvanceDecline																	
Period	Total Size	1-Variable			3-Variable						5-Variable						
		R-Square	Intercept	Milit	R-Square	Intercept	Fail	Milit	No	R-Square	Intercept	Exprsv	Fail	Milit	No	Pain	
1991	252	10%	0.7684	1.6113	15%	1.9840	-4.7257	1.7242	-13.8772	17%	0.5823	1.1697	-4.7438	1.6979	-13.0029	3.6746	
		(28.31)	(7.50)	(5.32)	(14.52)	(5.37)	(-2.58)	(5.77)	(-2.89)	(10.23)	(0.86)	(1.73)	(-2.55)	(5.46)	(-2.73)	(2.02)	
1992	253	2%	0.8488	8.6908	4%	0.3391	1.0240	8.7820	7.7336	6%	1.8978	-0.2284	0.9879	8.9869	2.2136	7.8708	
		(4.68)	(6.70)	(2.16)	(3.36)	(1.27)	(1.68)	(2.20)	(1.72)	(2.91)	(1.30)	(-1.44)	(1.63)	(2.26)	(1.60)	(1.76)	
1993	252	3%	0.7850	1.5852	5%	-1.8654	0.1855	1.4975	0.3407	7%	-1.5249	0.2002	1.4862	0.3861	-0.1919	-1.8861	
		(8.63)	(6.48)	(2.94)	(4.47)	(-1.48)	(1.48)	(2.78)	(1.39)	(3.49)	(-1.20)	(1.59)	(2.77)	(1.57)	(-1.36)	(-1.45)	
1994	251	3%	-1.1443	2.3697	6%	-1.5333	4.5374	2.5179	-5.1292	9%	-1.0280	1.2218	4.1021	-1.7631	2.5632	-4.8129	
		(7.52)	(-1.42)	(2.74)	(5.68)	(-1.86)	(2.65)	(2.89)	(-2.30)	(4.76)	(-1.16)	(1.65)	(2.38)	(-2.11)	(2.87)	(-2.17)	
1995	251	4%	-1.5092	1.2583	8%	-3.6449	0.4287	4.0605	1.1537	10%	-2.4409	-0.3599	0.4389	4.1479	-2.2767	1.2388	
		(9.68)	(-1.75)	(3.11)	(7.24)	(-3.08)	(1.61)	(2.81)	(2.87)	(5.41)	(-1.89)	(-1.80)	(1.64)	(2.87)	(-1.52)	(3.08)	
1996	253	4%	3.1085	-2.2121	7%	1.8930	10.8990	-2.6322	1.4669	8%	2.7234	-2.1443	11.4662	-2.5988	-2.6405	1.4398	
		(10.61)	(5.23)	(-3.26)	(5.99)	(2.22)	(2.04)	(-3.80)	(1.81)	(4.43)	(2.88)	(-1.39)	(2.16)	(-3.73)	(-1.52)	(1.78)	
1997	252	2%	4.3670	-0.2904	5%	4.1698	3.1737	6.5635	-0.3473	8%	5.4853	4.3221	-2.0252	8.0299	-3.9650	-0.3260	
		(4.49)	(2.97)	(-2.12)	(4.06)	(2.86)	(1.94)	(2.07)	(-2.53)	(4.44)	(3.45)	(2.61)	(-2.51)	(2.53)	(-2.24)	(-2.25)	
1998	251	2%	0.2720	2.7292	6%	-4.1906	3.9904	0.9484	4.1669	8%	-2.6735	4.0803	1.8072	0.8808	4.6305	-0.6509	
		(4.45)	(0.67)	(2.11)	(5.27)	(-2.83)	(2.74)	(2.66)	(2.58)	(4.13)	(-1.43)	(3.08)	(1.50)	(2.46)	(2.66)	(-1.68)	
1999	251	3%	0.6042	0.5847	5%	1.0140	0.4813	8.2235	-3.2508	9%	0.6674	0.4954	0.4468	-2.2073	5.7785	-3.0744	
		(6.65)	(3.91)	(2.58)	(4.77)	(3.16)	(2.11)	(1.96)	(-1.94)	(4.99)	(1.13)	(2.21)	(2.44)	(-2.85)	(1.37)	(-1.83)	
2000	251	3%	0.3128	1.8166	6%	-2.7051	0.5710	1.3381	0.7733	7%	-3.7099	-1.5032	0.1093	0.6584	1.6334	0.9889	
		(6.51)	(1.04)	(2.55)	(4.94)	(-2.40)	(2.25)	(1.84)	(1.71)	(3.73)	(-2.39)	(-1.30)	(1.43)	(2.55)	(2.11)	(2.00)	
2001	247	1%	1.5756	-1.2008	4%	1.3004	-2.7784	0.4800	-0.3831	6%	1.7971	-2.9809	0.6740	-0.4226	-0.4113	4.2678	
		(1.95)	(5.53)	(-1.40)	(3.72)	(2.81)	(-2.72)	(2.85)	(-1.91)	(2.97)	(2.04)	(-2.88)	(3.36)	(-1.30)	(-2.05)	(1.38)	
2002	251	2%	0.7305	12.3067	5%	2.0656	1.9997	11.8158	-0.6143	7%	5.4743	2.2454	11.9168	-0.7260	-0.6980	-0.3123	
		(5.95)	(3.72)	(2.44)	(4.17)	(1.93)	(2.26)	(2.35)	(-1.83)	(3.71)	(3.02)	(2.52)	(2.35)	(-2.05)	(-2.08)	(-1.53)	
EntirePeriod	3026	1%	0.9687	0.8064	2%	0.1536	0.7397	1.6272	0.5324	2%	0.6443	0.6963	0.6988	1.3488	-0.0565	0.5697	
		(37.78)	(31.53)	(6.15)	(19.26)	(0.78)	(5.55)	(3.45)	(2.98)	(13.60)	(1.82)	(2.93)	(5.10)	(2.79)	(-2.01)	(3.17)	

Table 7.7 Daily Stepwise Linear Regressions: YenUSDReturn = Categories, Categories Lag Markets by 1 Day

YenUSDReturn		1-Variable			3-Variable						5-Variable					
Period	Total Size	R-Square	Intercept	Persist	R-Square	Intercept	Econ@	Goal	Persist	R-Square	Intercept	Econ@	Goal	Milit	Persist	Virtue
1991	252	2% (4.16)	-0.0063 (-2.13)	0.0266 (2.04)	4% (3.75)	0.0015 (0.24)	-0.0017 (-1.97)	0.0135 (2.18)	0.0229 (1.76)	7% (3.53)	0.0182 (1.90)	-0.0024 (-2.61)	0.0125 (2.02)	-0.0042 (-1.95)	0.0273 (2.06)	-0.0070 (-1.90)
1992	254	3% (8.56)	0.0078 (2.92)	-0.0424 (-3.32)	6% (5.40)	0.0036 (1.18)	-0.0403 (-2.81)	0.0702 (1.89)	0.0825 (1.98)	8% (4.25)	0.0163 (1.92)	-0.0337 (-2.30)	-0.0724 (-1.56)	0.0715 (1.93)	-0.0029 (-1.48)	0.0885 (2.08)
1993	252	2% (5.31)	0.0127 (2.26)	-0.0045 (-2.30)	4% (3.89)	0.0206 (3.21)	-0.0046 (-2.38)	-0.0281 (-1.92)	-0.0240 (-1.48)	6% (3.06)	0.0179 (2.63)	-0.0088 (-1.30)	-0.0280 (-1.89)	-0.0252 (-1.91)	0.0324 (1.56)	0.0324 (1.43)
1994	253	2% (4.57)	0.0042 (1.95)	-0.0221 (-2.14)	6% (4.91)	-0.0004 (-0.05)	-0.0288 (-2.70)	-0.0043 (-1.44)	0.0120 (3.15)	7% (3.71)	0.0037 (0.39)	-0.0288 (-2.71)	-0.0050 (-1.63)	0.0214 (1.52)	-0.0040 (-1.13)	0.0117 (3.05)
1995	252	2% (14.97)	0.0125 (3.99)	-0.0178 (-3.87)	9% (8.20)	0.0199 (1.73)	0.0196 (2.59)	-0.0063 (-1.64)	-0.0160 (-3.44)	12% (6.46)	0.0075 (0.61)	0.0261 (3.32)	0.0092 (2.33)	-0.0100 (-2.44)	-0.0309 (-4.01)	0.0332 (1.51)
1996	253	3% (6.47)	-0.0111 (-2.44)	0.0018 (2.55)	6% (5.37)	-0.0098 (-1.66)	-0.0178 (-2.12)	0.0019 (2.68)	0.0283 (2.34)	8% (4.42)	-0.0117 (-1.10)	-0.0196 (-2.34)	0.0020 (2.76)	-0.0040 (-1.75)	0.0283 (2.11)	0.0033 (1.87)
1997	252	2% (4.96)	-0.0181 (-2.17)	0.0202 (2.23)	6% (5.40)	-0.0126 (-1.51)	-0.0288 (-1.83)	-0.0460 (-3.09)	0.0323 (3.35)	7% (4.00)	-0.0161 (-1.80)	-0.0331 (-2.08)	0.0104 (1.38)	-0.0539 (-3.42)	0.0330 (3.29)	-0.0451 (-1.45)
1998	252	5% (12.54)	0.0113 (3.21)	-0.0619 (-3.54)	8% (6.83)	0.0468 (3.52)	-0.0333 (-1.60)	-0.0035 (-2.15)	-0.0123 (-1.79)	9% (5.12)	0.0472 (3.51)	-0.0343 (-1.64)	-0.0032 (-2.01)	-0.1558 (-1.48)	-0.0125 (-1.82)	0.0186 (1.68)
1999	252	2% (5.26)	-0.0251 (-2.33)	0.0121 (2.29)	5% (4.56)	-0.0187 (-1.60)	-0.0164 (-1.95)	0.0636 (1.87)	0.0150 (2.78)	7% (3.79)	-0.0032 (-0.21)	-0.0156 (-1.84)	-0.0037 (-1.40)	0.0542 (1.55)	0.0147 (2.71)	-0.0609 (-1.81)
2000	253	3% (6.66)	0.0102 (2.67)	-0.0383 (-2.58)	6% (5.75)	0.0095 (1.64)	-0.0339 (-2.29)	0.0428 (2.77)	-0.0134 (-2.40)	9% (4.63)	0.0267 (2.70)	-0.0043 (-1.86)	-0.0343 (-2.32)	0.0414 (2.69)	-0.0145 (-2.59)	-0.0509 (-1.68)
2001	252	2% (5.58)	-0.0045 (-2.11)	0.0111 (2.36)	4% (3.55)	-0.0130 (-2.99)	0.0149 (2.97)	0.0140 (1.75)	0.0197 (1.38)	5% (2.85)	-0.0060 (-1.05)	0.0166 (3.20)	0.0139 (1.74)	0.0247 (1.68)	-0.0135 (-1.48)	-0.0157 (-1.58)
2002	252	3% (7.14)	0.0063 (2.48)	-0.0072 (-2.67)	5% (4.53)	0.0081 (2.69)	-0.0257 (-2.10)	-0.0090 (-2.86)	0.0180 (1.92)	8% (4.07)	0.0176 (3.32)	-0.0273 (-1.71)	-0.0310 (-2.51)	-0.0110 (-2.37)	0.0217 (3.21)	-0.0761 (-1.63)
EntirePeriod	3048	1% (6.59)	-0.0005 (-2.39)	0.0150 (2.57)	1% (5.12)	0.0001 (0.06)	0.0021 (2.23)	-0.0075 (-1.95)	0.0144 (2.45)	1% (3.77)	0.0004 (0.27)	0.0044 (1.34)	0.0019 (1.93)	-0.0082 (-2.03)	0.0141 (2.41)	-0.0020 (-1.68)

Table 7.8 Daily Stepwise Linear Regressions: OilReturn = Categories, Categories Lag Markets by 1 Day

OilReturn		1-Variable			3-Variable						5-Variable					
Period	Total Size	R-Square	Intercept	Decreases	R-Square	Intercept	Decreases	If	Weak	R-Square	Intercept	Decreases	EMOT	If	Need	Weak
1991	197	2% (4.20)	-0.0323 (-2.17)	0.1097 (2.05)	5% (3.74)	0.1269 (2.03)	0.1784 (3.01)	-0.0686 (-1.75)	-0.0673 (-2.04)	7% (2.89)	0.1184 (1.79)	0.1796 (2.97)	-0.1187 (-1.60)	-0.0692 (-1.75)	0.1281 (1.18)	-0.0569 (-1.69)
1992	199	4% (8.33)	-0.0233 (-2.90)	0.0173 (2.89)	9% (6.80)	0.0282 (1.54)	0.0156 (2.65)	-0.0445 (-2.57)	-0.0252 (-2.67)	12% (5.51)	0.0723 (2.70)	-0.2018 (-1.52)	0.0143 (2.41)	-0.0481 (-2.77)	-0.0282 (-2.97)	-0.0198 (-2.08)
1993	198	2% (4.66)	0.0136 (1.99)	-0.0437 (-2.16)	6% (3.78)	0.0370 (3.25)	-0.0517 (-2.53)	-0.0929 (-1.84)	-0.1287 (-1.74)	8% (3.13)	0.0335 (2.14)	0.0790 (1.64)	-0.0503 (-2.45)	-0.0692 (-1.37)	-0.0620 (-1.13)	-0.1195 (-1.62)
1994	198	3% (6.06)	-0.0515 (-2.39)	0.0339 (2.46)	7% (5.10)	-0.1506 (-3.84)	0.0192 (2.17)	0.0528 (2.04)	0.0300 (2.21)	11% (4.66)	-0.1176 (-2.86)	0.0170 (1.91)	0.0765 (2.81)	-0.0242 (-1.52)	0.0355 (2.62)	-0.1506 (-2.23)
1995	197	2% (3.16)	0.0183 (1.84)	-0.0136 (-1.78)	4% (2.71)	0.0073 (0.66)	0.2430 (1.69)	-0.0229 (-2.42)	0.0423 (1.59)	6% (2.58)	-0.0073 (-0.38)	-0.0540 (-1.83)	0.2540 (1.78)	-0.0231 (-2.43)	0.0371 (1.38)	0.0186 (1.67)
1996	195	4% (7.34)	-0.0771 (-2.75)	0.0870 (2.71)	9% (6.07)	0.0142 (0.26)	0.0890 (2.81)	-0.0292 (-2.08)	0.4546 (2.59)	13% (5.48)	0.0776 (1.36)	-0.1502 (-2.16)	0.0923 (2.92)	-0.0324 (-2.34)	-0.1706 (-2.06)	0.5073 (2.92)
1997	197	1% (1.61)	-0.0161 (-1.29)	0.0498 (1.27)	2% (1.47)	0.0196 (0.65)	0.0646 (1.60)	-0.0080 (-0.98)	-0.0680 (-1.56)	4% (1.39)	0.0285 (0.89)	0.1024 (1.99)	-0.0111 (-1.30)	-0.0787 (-1.78)	-0.0872 (-1.34)	0.0942 (1.04)
1998	198	2% (3.71)	-0.0410 (-1.91)	0.1984 (1.93)	5% (3.68)	-0.1379 (-2.08)	-0.1427 (-2.08)	0.1951 (1.92)	0.0693 (2.22)	8% (3.52)	-0.2207 (-2.93)	-0.2105 (-2.87)	0.0250 (1.86)	0.2117 (2.10)	0.6011 (1.78)	0.0778 (2.51)
1999	198	2% (4.23)	0.0235 (2.38)	-0.0450 (-2.06)	7% (4.96)	-0.0924 (-1.73)	0.0103 (1.69)	-0.0704 (-3.07)	0.1859 (2.39)	10% (4.29)	-0.0208 (-0.35)	0.0122 (2.00)	-0.0714 (-3.13)	0.2227 (2.83)	-0.0604 (-1.73)	-0.0714 (-2.05)
2000	198	3% (5.27)	0.0730 (2.33)	-0.0273 (-2.30)	5% (3.57)	0.0639 (1.76)	-0.0545 (-1.59)	-0.0316 (-2.64)	0.1097 (2.09)	7% (3.05)	0.0060 (0.13)	-0.0468 (-1.36)	0.0563 (1.64)	-0.0279 (-2.32)	0.1000 (2.81)	0.2500 (1.57)
2001	196	3% (6.77)	0.0274 (2.63)	-0.1721 (-2.60)	8% (5.51)	0.1140 (3.11)	-0.1845 (-2.80)	-0.7199 (-2.20)	-0.0432 (-2.07)	10% (4.45)	0.1613 (3.87)	-0.1689 (-2.48)	-0.0593 (-1.42)	-0.6401 (-1.96)	-0.1813 (-1.79)	-0.0420 (-2.04)
2002	196	1% (2.55)	-0.0470 (-1.53)	0.0149 (1.60)	5% (3.19)	-0.0645 (-1.96)	-0.0394 (-1.97)	0.0168 (1.62)	0.0121 (2.15)	8% (3.10)	-0.0600 (-1.22)	-0.0441 (-2.22)	0.0201 (2.16)	0.0136 (2.41)	-0.0129 (-1.66)	0.0501 (1.76)
EntirePeriod	2378	0% (3.36)	-0.0249 (-1.80)	0.0028 (1.84)	0% (3.65)	-0.0146 (-0.93)	0.0046 (2.77)	-0.0321 (-1.97)	0.0052 (-1.74)	1% (2.82)	-0.0163 (-1.03)	0.0049 (2.88)	0.0111 (1.12)	-0.0200 (-1.54)	-0.0341 (-1.98)	-0.0042 (-1.40)

Table 7.11 Daily Stepwise Linear Regressions: NYSEAdvanceDecline = Categories, Categories Lag Markets by 1 Day

NYSEAdvanceDecline																
Period	Total Size	1-Variable			3-Variable					5-Variable						
1991	196	R-Square 9%	Intercept 0.8117 (20.19)	Milit 1.4814 (4.49)	R-Square 12%	Intercept 0.1825 (8.97)	Increas -1.9059 (-1.93)	Milit 1.5319 (4.83)	Virtue 0.9822 (1.59)	R-Square 15%	Intercept -0.2769 (-0.19)	Increas -2.2940 (-2.21)	Milit 1.7066 (4.54)	Ovrst 0.8262 (1.52)	SureLw -2.9861 (-2.17)	Virtue 1.2028 (1.91)
1992	199	R-Square 3%	Intercept 0.7967 (6.31)	Milit 1.8162 (4.71)	R-Square 8%	Intercept 3.9259 (5.37)	Active -0.4355 (-2.45)	EMOT 2.1388 (2.02)	Milit 1.6937 (2.38)	R-Square 12%	Intercept 4.4780 (5.49)	Active -0.3847 (-2.18)	Arousal -5.1928 (-2.86)	EMOT 3.2523 (2.30)	Fail -2.6547 (-1.81)	Milit 1.6736 (2.39)
1993	199	R-Square 3%	Intercept 1.8767 (5.71)	Compleat -2.4016 (-2.39)	R-Square 7%	Intercept 1.2944 (4.75)	Compleat -2.6087 (-2.61)	Decreas 1.2475 (1.66)	Yes 4.9234 (2.54)	R-Square 10%	Intercept 0.3891 (0.73)	Compleat -2.8056 (-2.83)	Decreas 1.9023 (2.41)	RspTot 1.1573 (1.78)	Try 2.1503 (1.77)	Yes 4.5219 (2.34)
1994	198	R-Square 2%	Intercept 1.9550 (4.76)	Polit@ -0.4659 (-2.18)	R-Square 5%	Intercept 2.2840 (3.13)	Fail 3.1246 (3.77)	Polit@ -0.4979 (-1.42)	Vice -1.0155 (-2.34)	R-Square 7%	Intercept 0.8892 (2.89)	Arousal 3.1832 (1.02)	Fail 4.0184 (1.67)	Polit@ -0.5527 (-1.81)	Try 4.9661 (1.71)	Vice -1.0825 (-1.97)
1995	197	R-Square 2%	Intercept -0.8654 (4.28)	Undrst 0.9661 (-0.87)	R-Square 5%	Intercept -1.7616 (3.49)	Legal -0.5333 (-1.33)	Passive 0.5259 (1.74)	Undrst 0.9963 (2.15)	R-Square 7%	Intercept -2.2807 (-1.69)	Compleat 1.8389 (1.38)	Legal -0.5368 (-1.91)	Passive 0.4806 (1.51)	Undrst 0.9407 (2.03)	WbPsyc 2.1367 (1.16)
1996	197	R-Square 3%	Intercept 1.0007 (5.35)	No 5.3863 (11.10)	R-Square 7%	Intercept -0.2679 (5.15)	Exprsv 1.1300 (-0.65)	Goal 1.8305 (2.42)	No 5.2525 (2.27)	R-Square 9%	Intercept 0.6217 (3.96)	Exprsv 1.0094 (2.14)	Goal 1.6557 (2.03)	If -1.4478 (-1.85)	Need 2.0687 (1.35)	No 5.5942 (2.42)
1997	197	R-Square 3%	Intercept -0.5355 (5.19)	Weak 1.1583 (-0.69)	R-Square 7%	Intercept 2.0775 (4.56)	Active -0.4878 (-2.39)	Virtue 0.9532 (1.76)	Weak 1.3168 (2.61)	R-Square 9%	Intercept 1.5500 (3.59)	Active -0.4233 (-2.00)	Need -2.4608 (-1.50)	Think 7.3156 (1.40)	Virtue 1.1728 (2.11)	Weak 1.3487 (2.68)
1998	198	R-Square 2%	Intercept 1.8547 (3.24)	Legal -0.6027 (-1.80)	R-Square 4%	Intercept 0.7406 (2.36)	Exprsv 0.8362 (1.07)	Legal -0.5209 (-1.55)	Pain 2.5584 (1.45)	R-Square 6%	Intercept -0.9998 (-2.34)	Exprsv 1.0229 (-0.70)	Legal -0.6783 (-1.77)	Pain 4.3717 (1.98)	Positiv 0.4690 (2.24)	WbPsyc -4.1603 (1.52)
1999	199	R-Square 5%	Intercept -1.1071 (10.97)	Means 0.7311 (3.31)	R-Square 7%	Intercept -1.2422 (5.02)	Means 0.5949 (-0.92)	Undrst 0.6448 (2.55)	Virtue -0.5002 (-1.29)	R-Square 8%	Intercept -0.9453 (-3.59)	Compleat 2.0237 (-0.59)	Means 0.5889 (1.58)	Positiv -0.3650 (-2.53)	Undrst 0.6867 (1.54)	Virtue -0.1736 (-0.37)
2000	198	R-Square 3%	Intercept 0.1236 (6.93)	Need 3.6929 (0.34)	R-Square 9%	Intercept -0.4817 (6.34)	Compleat -2.8721 (-2.19)	Need 3.4916 (2.53)	Weak 1.0634 (2.91)	R-Square 11%	Intercept 0.6107 (4.74)	Compleat -2.9549 (-0.78)	Feel -9.0640 (-2.25)	Need 3.4399 (-1.57)	Vice 0.8311 (2.51)	Weak 0.9546 (2.53)
2001	192	R-Square 2%	Intercept 2.6682 (3.95)	SureLw -1.6594 (-1.99)	R-Square 7%	Intercept 2.1287 (4.80)	Milit 1.0089 (2.80)	No 12.0311 (2.37)	SureLw -1.7645 (-2.15)	R-Square 10%	Intercept 2.0835 (4.21)	EMOT -2.1348 (-2.65)	Milit 1.3196 (-1.95)	No 10.6767 (3.20)	SureLw -1.2432 (2.11)	Think 10.8550 (-1.47)
2002	196	R-Square 3%	Intercept 2.0731 (5.49)	Pain -3.6933 (-2.34)	R-Square 8%	Intercept 2.2623 (5.50)	Compleat 2.8612 (3.22)	Legal -0.9694 (-2.88)	Pain -4.4430 (-2.82)	R-Square 11%	Intercept 2.6144 (4.55)	Arousal -4.2069 (-3.03)	Compleat 2.3867 (-1.95)	EMOT 3.2572 (1.78)	Legal -1.1162 (-1.78)	Pain -6.5204 (-2.98)
EntirePeriod	2377	R-Square 1%	Intercept 0.9774 (28.83)	Milit 0.7881 (5.37)	R-Square 2%	Intercept 0.5383 (15.07)	Milit 0.7978 (3.70)	Vice -0.3866 (-2.65)	Weak 0.4493 (3.99)	R-Square 2%	Intercept 0.6828 (10.09)	Compleat -0.6257 (-1.56)	Milit 0.7980 (5.16)	Think 2.7746 (1.68)	Vice -0.4033 (-2.77)	Weak 0.4547 (4.02)

4.3 Event Study Results

The results of the event study can be broken down into three types based on the category score averages relative to the entire period averages before, during, and after important events occur:

- **Big Jumps:** the category score shows a significantly large jump at some point after events of a particular type occur.
- **Trends:** the category score shows a significant and continued buildup or decline at some point after events of a particular type occur.
- **Nothing:** no jumps or trends are present, but the average category score deviates significantly from entire period average around event occurrences.

Figures 8, 9, and 10 show the three result types Big Jumps, Trends, and Nothing described above, respectively. Each plot has on the horizontal axis the event day, which is the number of days before or after the day of the event. For example day 0 is the day closest to the date of the events, day -1 is the day closest to one day before the events, and day 1 is the day closest to one day after the events. The vertical axis is the event excess average as previously discussed in the event studies methodology section. The event excess average is the percentage difference between the average category score on the event day and the average category score over the entire period of 1991-2002. The title of each plot is formatted as Event Type / Category.

Figure 8 Event Study Results: Big Jumps

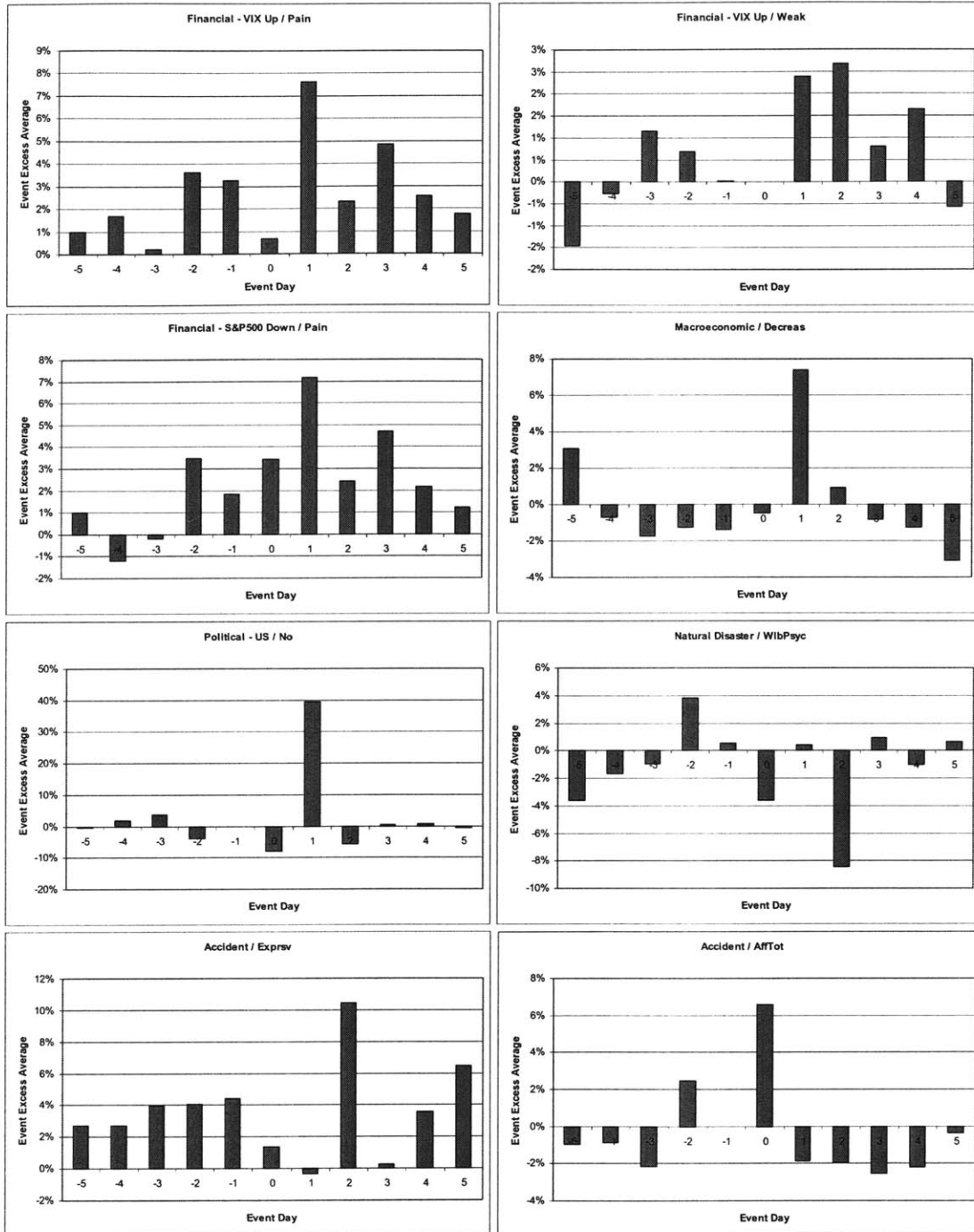


Figure 9 Event Study Results: Trends

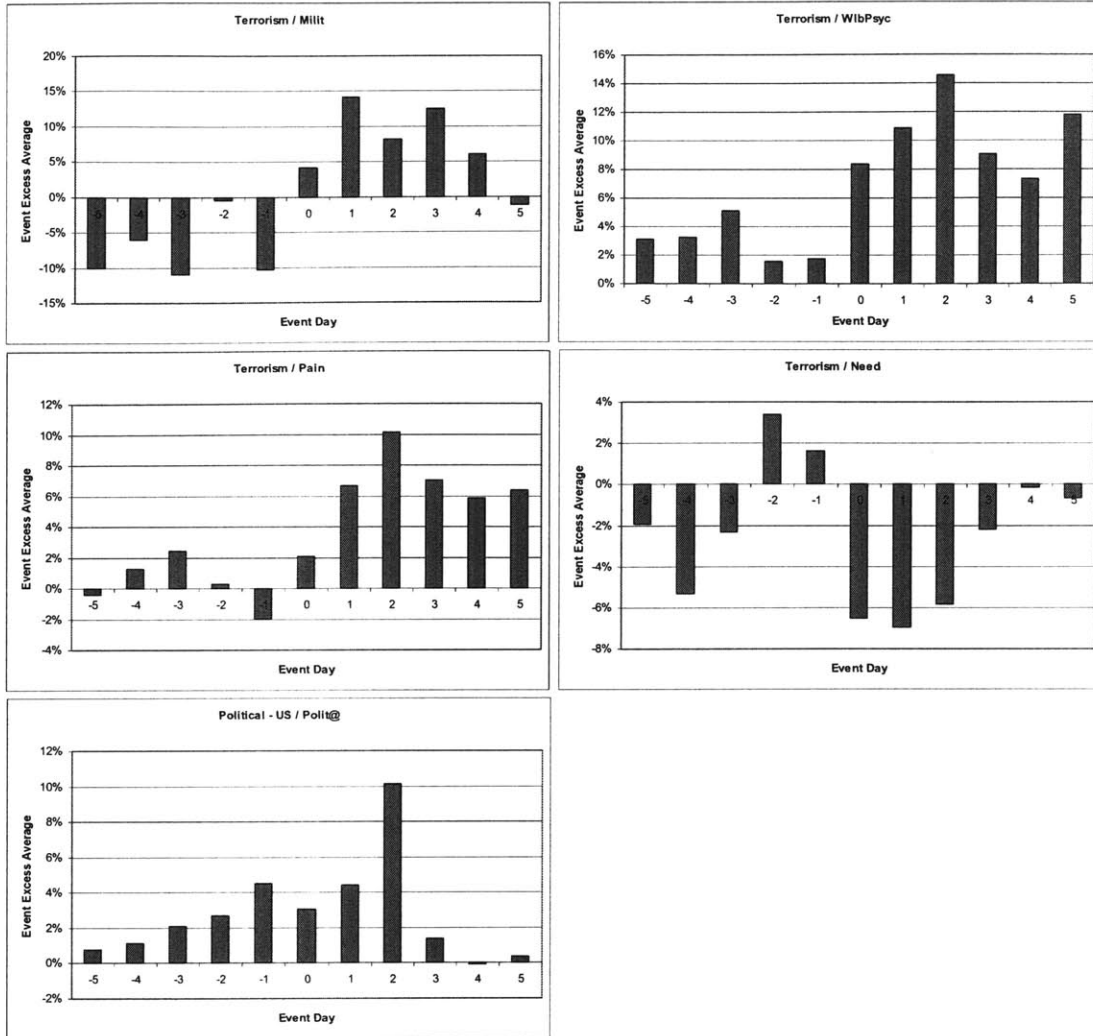
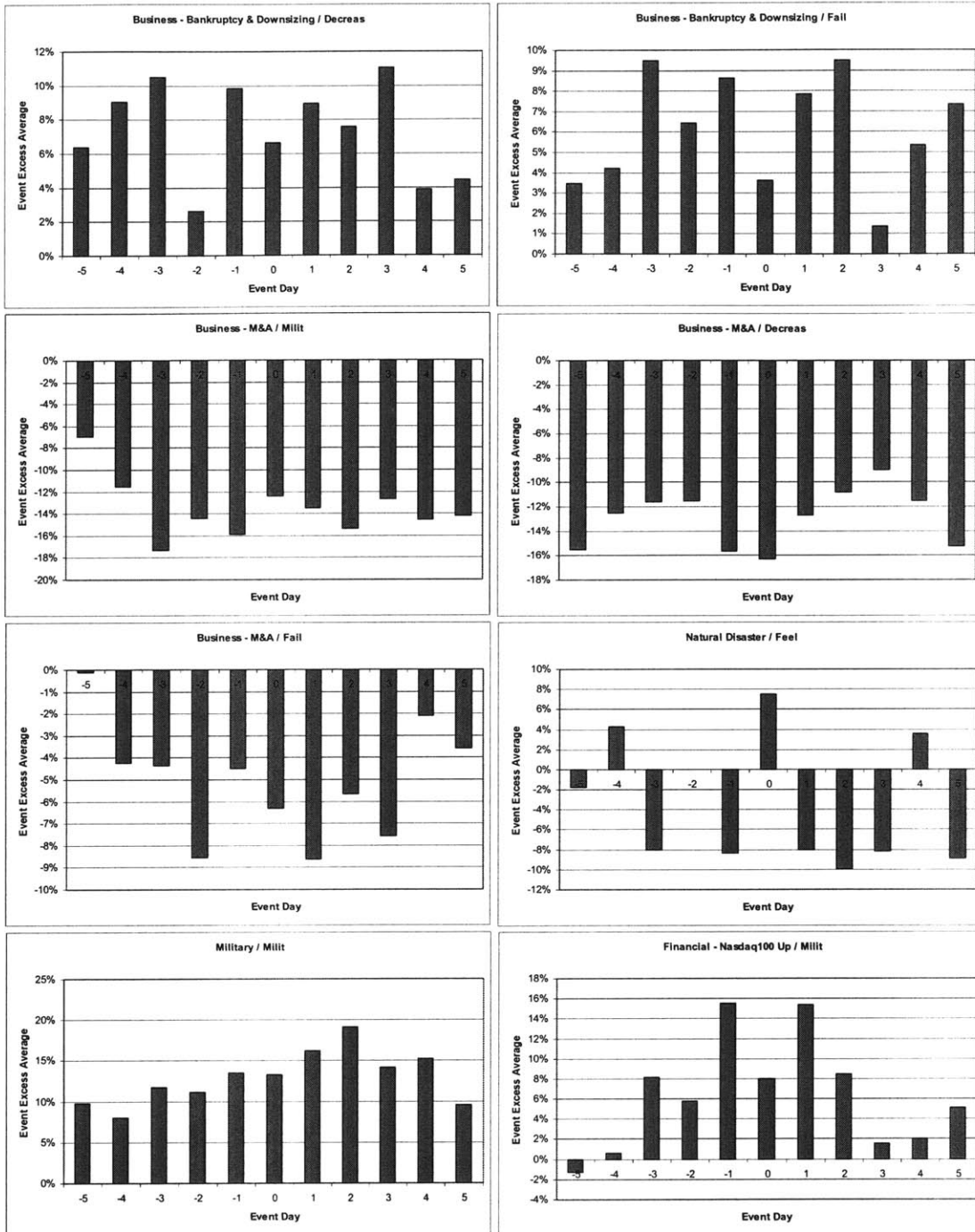


Figure 10 Event Study Results: Nothing



5. Discussion

5.1 Correlations

Using the top five most strongly correlated categories with each market variable from 1991-2002 from the results section, we compiled frequency counts of how many times each category showed up in the top five across all market variables. The frequency counts are compiled across all years 1991-2002, all market variables, and all three lead/lag types (no lag, markets lag categories by 1 day, and categories lag markets by 1 day). Table 8 shows the result for the top ten such most frequently strongly correlated categories. We note that four of the top five most frequently strongly correlated categories in the table are emotional categories with negative connotations: Weak, Negativ, Pain, and Fail.

Table 8 Frequency Counts of the Top 5 Most Strongly Correlated Categories
(Across all years 1991-2002, all market variables, all 3 lag types)

Category	Top 5 Count
Weak	16
Positiv	11
Negativ	10
Pain	9
Fail	8
Comple	7
Decreas	7
Active	6
Need	6
Means	5

Correlations often vary greatly from year to year for the same category and market variable pairs. In other words, a category that is strongly correlated with a particular market variable during one year may not be correlated with the same market variable at all the next year. The reason behind this fluctuation in correlations may simply be that no one set of emotions consistently dominates across the time period we've examined. Likewise, no particular years exhibit consistently strong correlations across market variables or categories.

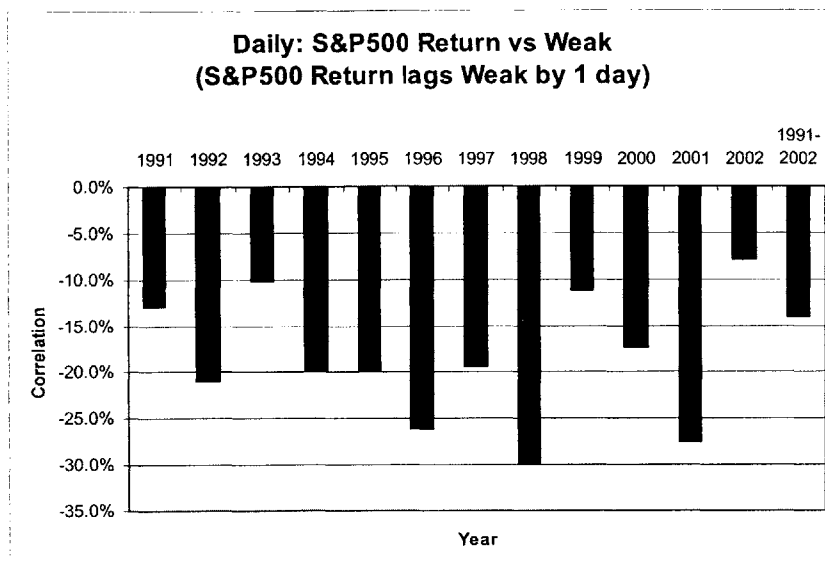
Another noteworthy result is that for daily correlations, markets lagging categories correlate slightly better than categories lagging markets. This could mean that markets are a better predictor of emotions than emotions are of markets. Contemporaneous or no lag correlations are generally weaker than both lagged cases. One should consider that even in the no lag case, there is an inherent lag since the Wall Street Journal is published

in the morning and market returns on the same day should reflect the information in that morning's publication.

From a purely strength of correlation standpoint, market variables like CBOE Put/Call and S&P500 Return Square show stronger correlations with the category scores than other market variables such as returns. This implies that emotions correlate better with market volatility than with market returns. We also note that the CBOE Put/Call and S&P500 Return Square are two of the most stationary market variables across the time period of 1991 to 2002. It is possible that correlations between the category scores and these two market variables are stronger partly due to the higher stationarity of the market variables, given that the category scores are also highly stationary.

We found it often difficult to interpret the meaning of many of the correlations. For example, if a category like Need is strongly and positively correlated with S&P500 Return in year 2000 (26% correlation coefficient with P-value of 0.0002), what does that mean? If we can't interpret that result using our current understanding of emotions and financial markets, does that mean we should ignore this result? More work may need to be done to define a framework for determining what set of characteristics each result must exhibit in order to be considered important. For example, one may argue that if a category consistently exhibits strong correlations with a market variable across all the years and that the correlation coefficients are of the same sign across these years, then it warrants attention. If one were to follow such a framework, the daily correlations between the Weak category score and S&P500 Return for the case where Weak lags S&P500 Return by 1 day, as shown in Figure 11, may be considered important. The corresponding correlation coefficients and associated P-values used in the plots are given below the plot.

Figure 11 Daily Correlation: S&P500 Return vs Weak, S&P500 Return lags Weak by 1 day



	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	1991-2002
Corr	-12.9%	-20.9%	-10.2%	-19.9%	-19.8%	-26.1%	-19.4%	-29.8%	-11.1%	-17.4%	-27.6%	-7.9%	-14.1%
P-val	7.0E-02	2.9E-03	1.5E-01	4.8E-03	5.0E-03	2.0E-04	6.0E-03	1.9E-05	1.2E-01	1.4E-02	9.7E-05	2.7E-01	5.0E-12

The next question to ask once we've selected the correlations we want to examine is "what is driving the correlations?" One reasonable hypothesis is that the outliers, or the most extreme subset of a category's time series, drive correlations. To investigate the validity of this hypothesis, we replaced increasing percentages of non-outlier category values with randomly generated values from a normal distribution with the same mean and standard deviation as the non-outliers and observed how that affects the strengths of the correlations. To calculate the outliers of a time series S , we first subtract out the mean μ_S to get \bar{S} . We then take the absolute value of the resulting time series to get $|\bar{S}|$. The outliers are all elements whose absolute value is greater than the n^{th} percentile of this new time series given by $Percentile(|\bar{S}|, n)$.

$$\bar{S} = S - \mu_S$$

$$p = Percentile(|\bar{S}|, n)$$

$$S_{outlier} = \{s \in S \mid |s| > p\}$$

Once we've computed the outliers, we then replace the rest of the sample, the non-outliers $S_{nonoutlier}$, with randomly generated samples from a normal distribution with the same mean and standard deviation S_{random} , to observe how that affects correlations.

$$S_{nonoutlier} = \{s \in S \mid s \notin S_{outlier}\}$$

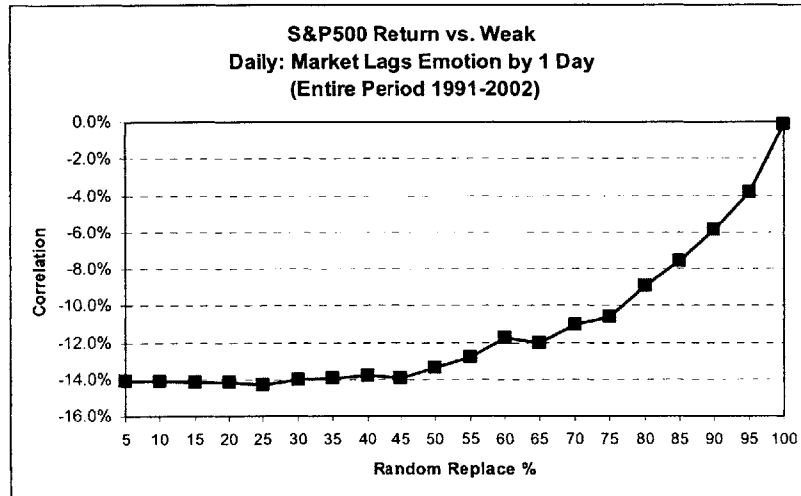
$$\mu = \text{mean}(S_{nonoutlier})$$

$$\sigma = \text{stdev}(S_{nonoutlier})$$

$$N(\mu, \sigma) \rightarrow S_{random}$$

Let's take the same Weak and S&P500 Return correlation example discussed previously where S&P500 Return lags Weak by 1 day. Figure 12 shows the change in daily correlation between Weak and S&P500 Return as increasing percentages (increasing n in the *Percentile* formula above) of the category's non-outliers were replaced by randomly generated values. The correlation and random sample generations were performed over the entire period of 1991-2002. The result shown was averaged over 20 independent random generation and correlation analysis trials. As the plot in Figure 12 reveals, we were able to replace a little more than half of the non-outlier Weak category score values without affecting the correlation. The most extreme 5% of the Weak category score values accounted for almost a third of the correlation.

Figure 12 Impact of Outliers on Correlations: S&P500 Return vs. Weak (1991-2002)



5.2 Regressions

The linear regression results are related to correlation results in the sense that the n categories that are most strongly correlated with a market variable will also be in the corresponding best n -variable linear regression model. Below we select a few Market = Categories daily linear regression models from the results section that have particularly high and significant R-Square values. Table 9 shows a few such significant models for the no lag case. Table 10 shows a few models where categories lag markets by 1 day.

Unsurprisingly, the market variables that showed strong correlations, CBOE Put/Call and S&P500 Return Square, appeared again in the significant regression models.

Table 9 Significant Daily Linear Regression Models: Market = Categories, No Lag

Market Variable	Period	Total Size	1-Variable			3-Variable					5-Variable						
			R-Square	Intercept	Weak	R-Square	Intercept	Comple	Think	Weak	R-Square	Intercept	Comple	Pain	Think	Weak	Yes
CBOEPutCall	1998	251	13%	-0.1522	0.5397	22%	0.4620	-1.3471	-3.0368	0.4585	25%	0.4328	-1.2312	0.9337	-3.2198	0.3532	-0.3391
			(38.78)	(-1.14)	(6.23)	(23.54)	(2.66)	(-4.89)	(-2.45)	(5.44)	(16.68)	(2.52)	(-4.49)	(2.51)	(-2.62)	(3.77)	(-2.22)
CBOEPutCall	2000	251	R-Square	Intercept	Decreases	R-Square	Intercept	Decreases	Econ@	Positiv	R-Square	Intercept	Decreases	Econ@	EMOT	Passive	Positiv
			17%	0.2906	1.5409	33%	1.8615	1.4668	-0.0801	-0.2568	38%	1.2603	1.1115	-0.0631	0.5006	0.1913	-0.2931
SP500ReturnSquare	1998	251	9%	0.0036	-0.0008	18%	0.0096	0.0016	-0.0007	0.0007	21%	0.0014	0.0018	-0.0005	-0.0008	0.0013	0.0009
			(24.43)	(5.19)	(-4.94)	(18.03)	(0.74)	(3.71)	(-4.44)	(2.86)	(13.01)	(1.44)	(4.17)	(-1.96)	(-4.99)	(2.36)	(3.35)
SP500ReturnSquare	EntirePeriod	3028	R-Square	Intercept	Legal	R-Square	Intercept	Legal	Negativ	Positiv	R-Square	Intercept	Legal	Negativ	Pain	Positiv	Undrst
			5%	0.0005	-0.0003	8%	0.0007	-0.0003	0.0001	-0.0002	8%	0.0010	-0.0002	0.0001	0.0008	-0.0002	-0.0002
			(161.10)	(16.05)	(-12.69)	(83.14)	(5.71)	(-12.22)	(8.02)	(-5.49)	(55.07)	(6.07)	(-8.37)	(4.03)	(3.93)	(-5.61)	(-3.30)

Table 10 Significant Daily Linear Regression Models: Market = Categories, Categories Lag Markets by 1 Day

Market Variable	Period	Total Size	1-Variable			3-Variable					5-Variable						
			R-Square	Intercept	Weak	R-Square	Intercept	Comple	Ovrst	Weak	R-Square	Intercept	Comple	Ovrst	Try	Weak	Yes
CBOEPutCall	1998	198	10%	-0.0111	0.4491	18%	1.2561	-0.9456	-0.2329	0.3544	20%	1.1930	-0.9538	-0.2567	0.9395	0.3689	-0.2847
			(22.19)	(-0.08)	(4.71)	(13.80)	(3.40)	(-3.10)	(-2.75)	(3.75)	(9.80)	(3.21)	(-3.16)	(-3.05)	(1.88)	(3.94)	(-1.82)
CBOEPutCall	2000	198	R-Square	Intercept	Decreases	R-Square	Intercept	Decreases	Increases	Positiv	R-Square	Intercept	Decreases	Increases	Positiv	Vice	
			12%	0.3414	1.2625	23%	1.4430	1.1800	-0.5692	-0.1911	26%	1.3510	1.1516	-0.5063	0.0629	-0.1862	-0.2401
NYSEVolumeReturn	2002	196	7%	-0.6249	0.6383	17%	-0.1198	-0.4919	-0.7886	0.6360	23%	0.1684	-0.6653	0.9530	-0.7305	0.5431	-0.1722
			(14.19)	(-3.57)	(3.77)	(13.00)	(-0.80)	(-3.52)	(-3.70)	(3.84)	(11.45)	(0.54)	(-4.68)	(3.20)	(-3.58)	(3.35)	(-2.25)
SP500ReturnSquare	EntirePeriod	2377	R-Square	Intercept	Legal	R-Square	Intercept	Legal	Negativ	Positiv	R-Square	Intercept	Legal	Negativ	Positiv	Undrst	
			5%	0.0005	-0.0003	7%	0.0006	-0.0003	0.0001	-0.0001	8%	0.0009	0.0002	-0.0002	0.0001	-0.0001	-0.0002
			(121.58)	(14.22)	(-11.03)	(57.95)	(4.55)	(-10.58)	(6.39)	(-3.83)	(39.96)	(5.85)	(2.76)	(-6.08)	(3.64)	(-4.72)	(-4.53)

However, challenges similar to those we encountered in interpreting and analyzing the correlations results resurface in the regressions. The fact that the Yes category appeared in significant 5-variable regression models of CBOEPutCall = Categories in 1998 for both lag and no lag cases could cause concern. The Yes category composes only of 20 words indicating agreement such as “yeah”, “yes”, “sure”, and “right”. Because of its undesirable statistical properties, including non-stationarity and high kurtosis and skewness, the Yes category can be considered a noise category. Therefore, one can interpret the appearance of Yes in a significant regression model as a mere coincidence. But at the same time, it is hard to assume that the Yes category cannot possibly be related to the CBOE Put/Call ratio, given its reasonably significant parameter T-value of -2.2 in the 1998 no lag 5-variable model.

5.3 Event Study

One challenge associated with the event study analysis is in defining event dates. In other words, it is hard to determine the specific date before which the public has little or no knowledge of the event. We would like to choose dates on which the events first became known to the public and therefore would have the largest impact on the category scores. Take for example military and business events. The day a war is officially declared is probably not the day that the military engagement first became known to or highly anticipated by the public. Likewise, mergers and acquisitions are public knowledge before the deals themselves close. With the challenge of event date selection in mind, we will now examine the event study results.

Amongst the Big Jumps event study results, we see that significant increases in the VIX cause the Pain and Weak category scores to increase suddenly. A significant down movement in the S&P500 also causes a sharp increase in the Pain category score. These results seem to point again to negative emotions and the prominence of their relationships to financial events as we've encountered previously in the correlation results. However, the impact of macroeconomic events on the Decrease category score is harder to interpret since not all macroeconomic events we've selected are negative.

In the Trends event study results set, it isn't surprising to see that terrorism events have a sustained effect on Pain, WlbPsyc, and Milit category scores, causing these scores to increase immediately after the event dates and remain at fairly high levels afterwards. Terrorism event dates are fairly easy to select since it is very unlikely for the public to foresee terrorist acts before they occur, so the date a terrorist event actually occurred is the event date. The dominant presence of negative events in both the Big Jumps and Trends category suggests that there is a stronger relationship between negative events and negative emotions than between positive events and positive emotions.

In the Nothing event study results set, we find event types such as business and military that do not usually have clear event dates where the public knows nothing of the event before the event date. Therefore, we see cases such as the Milit category scores staying at a significantly high level before, during, and after military event dates. Likewise, the Fail category scores are consistently high throughout the period before, during, and after bankruptcy and downsizing event dates.

6. Future Work

6.1 Emotional Category Creation

The analysis presented so far assumes that the category scores are computed according to a reliable dictionary of word categories. We assume that the categories used by General Inquirer accurately categorize words according to their meanings. But the process of categorizing words into subject and psychological categories is a highly subjective procedure. A word may fall under many different categories according to the most obvious and least obvious meanings, not to mention that the subtle meaning of certain words change over time. Words themselves may go into and out of fashion over time. Another obvious problem is that there may not be enough categories to fully capture the sentiment on a given set of days. For example, one category that is lacking from the General Inquirer is a general fear category.

One approach to address the many shortcomings of relying on a given dictionary of categories is to build our own categories. If we start with a set of days we know are days where certain emotions dominate, we can observe through statistical filtering what set of words occur more often than usual on those days. One can think of this approach as a backward approach where we start with a corpus of training text and arrive at a set of categories. A detailed discussion of this approach can be found in Appendix B. Of course, such an approach has its own share of challenges. Picking days on which certain emotions dominate is highly subjective as well and careful statistical procedures must be applied to ensure that noise words are not confused with significant words and vice versa. Moreover, we must exercise care as to not pick days that are too extreme like September 12th, 2001 so that we don't train the system on outliers.

6.2 Phrase-Level Textual Processing

The success of the Emotional News project is directly dependent on the accuracy with which emotional information can be extracted from plain text news article. To date, natural language processing is still a developing field. It is an inherent challenge for computers to recognize the emotional content in the natural language of news articles.

The effectiveness of word-level analysis used for this project can be greatly improved by adding phrase-level analysis. The goal is to capture meanings that span multiple words and that cannot be captured by analyzing the words in isolation. A simple example is a two-word phrase starting with the word “not”, like “not urgent”. With a word-level analysis system such as the General Inquirer, “not” is recorded separately from “urgent” so that “urgent” still triggers an increase in the category score of Strng (Strong). A phrase-level system should recognize that in “not urgent”, the word “not” negates the effect of the word “urgent”.

A new phrase-level textual analysis approach developed by Liu, Hugo, Lieberman, and Selker (2003) at the MIT Media Lab uses large-scale real-world knowledge about the inherent affective nature of everyday situations (such as “getting into a car accident”) to classify sentences into basic emotion categories. Open Mind Commonsense was used as a real world corpus of 400,000 facts about the everyday world. Four linguistic models are combined for robustness as a society of commonsense-based affect recognition. These models cooperate and compete to classify the affect of text.

Results from testing the phrase-level system in an email writing application suggest that the approach is robust enough to enable plausible affective text user interfaces. While the phrase-level analysis system may not be as thorough in analyzing sophisticated text as the word-level system, it is more reliable at recognizing the emotional content of basic phrases, which may account for a sizable portion of news articles.

To integrate a phrase-level system with the current word-level system, we can pass the input text corpus into the phrase-level system first and then pass the unanalyzed text into the word-level system. An algorithm can then be applied to coalesce the different sets of scores produced by the phrase-level and word-level systems.

6.3 Subject-Specific News

Analyzing all news articles provides results on a macro level and forms a good foundation for measuring the validity of the general methodology presented here. However, using subject-specific news such as event-specific or security-specific news might provide more concrete and actionable results.

Event-specific news concerns particular major events such as a war, a presidential election, a major company bankruptcy, etc. Emotional scores extracted from such news can be analyzed in conjunction with movements in major market indices or in particular securities. Event-specific news can be obtained from a keyword search of news articles or from dedicated columns in news sources (i.e. “War on Iraq” column).

Security-specific news includes earnings announcements, corporate press releases, and other information concerning a specific company and its stock. Such news can be analyzed against changes in the respective security’s price and volume as well as changes in market indices and macroeconomic indicators. Security-specific news can be obtained through a keyword search of news articles or from a financial news source such as Yahoo! Finance.

The same textual processing architecture and statistical methodologies can be used to discover relationships between event-specific and security-specific news and financial markets. Some modifications may be necessary to handle smaller bodies of texts so that computed scores, which depend on total word counts, are not skewed.

6.4 Emotional Index

To visualize and distribute the results of Emotional News, an index can be calculated to reflect how the levels of different extracted emotional and subject category scores change with time. Such an “emotional index” can be dynamically compared to changes in market or security indices. To generate a web-based visual representation of the index, a web-enabled component can be built to translate numerical results of textual processing and statistical analysis into appropriate graphics. One instance of this graphical representation can be a web-accessible “market weather forecast” so that high levels of negative emotion scores in conjunction with a low level of positive scores would translate into a “cloudy” image, for example.

7. Conclusion

We've presented here a first step towards developing a quantitative model that relates investor emotions to financial markets. First, we measured investor emotions quantitatively on a "macro" level through content analysis of daily Wall Street Journal articles. The output of this step is a daily set of General Inquirer subject and emotional category scores that represent the percentage of words found in each category relative to the total number of words in the Wall Street Journal on that day. The statistical properties of these category scores are favorable in that they show strong stationarity and that many category score distributions are sufficiently close to a normal distribution. We showed that subject category scores such as Milit successfully picked up on the appropriate military events from the news text. A key assumption we made before going forward is that emotional-type content can be extracted from the text in the same manner and as successfully as subject-type content.

Next, we ran daily correlations and regressions between the category scores and broad market indices variables such as return, volume, and volatility to determine whether there is a relationship. We found that negative emotions are more strongly correlated with market variables than positive emotions. We also found that markets lagging emotions correlated better than emotions lagging markets, or in other words, markets are a better predictor of emotions than emotions of markets. There also appears to be a stronger relationship between emotions and market volatility than with market returns. From year to year, however, correlations generally fluctuated greatly such that a category that correlated strongly with a market variable during one year may not be correlated at all with the same market variable the following year or may have a correlation coefficient of the opposite sign. In investigating the source of the correlations, we found that category score outliers, the most extreme values, often drive the correlations. Regression results reaffirm the correlation results, with a few models showing significantly large R-Square's. Event study results again show that negative emotions are more strongly related to negative important events than positive emotions are to positive events. However, there are inherent difficulties in defining clear event dates for some event types.

A challenge we encountered that remains to be fully addressed is how to interpret the results of the correlation and regressions. More specifically, how do we extrapolate regression models such as $SP500ReturnSquare = \alpha + \beta_1 \cdot If + \beta_2 \cdot Positiv + \beta_3 \cdot Weak$ to high-level concepts? And if we encounter statistically significant results that are not consistent with our understanding of emotions and financial markets, what additional framework should we use to reconcile our existing understanding with the conflicting results?

The approach presented here will hopefully generate some future work in using quantitative approaches to capture investor sentiment so that concrete relationships

between investor emotions and financial markets can be discovered. The methodologies employed here leave much room for improvement. First, the current word-level textual analysis can be extended to a phrase-level approach to better extract content from plain text. Second, a different set of emotional categories can be developed by analyzing text on days known to be dominated by certain emotions. Lastly, there are plenty of opportunities for dissecting subject or security specific news and for visualizing the results of emotion extraction.

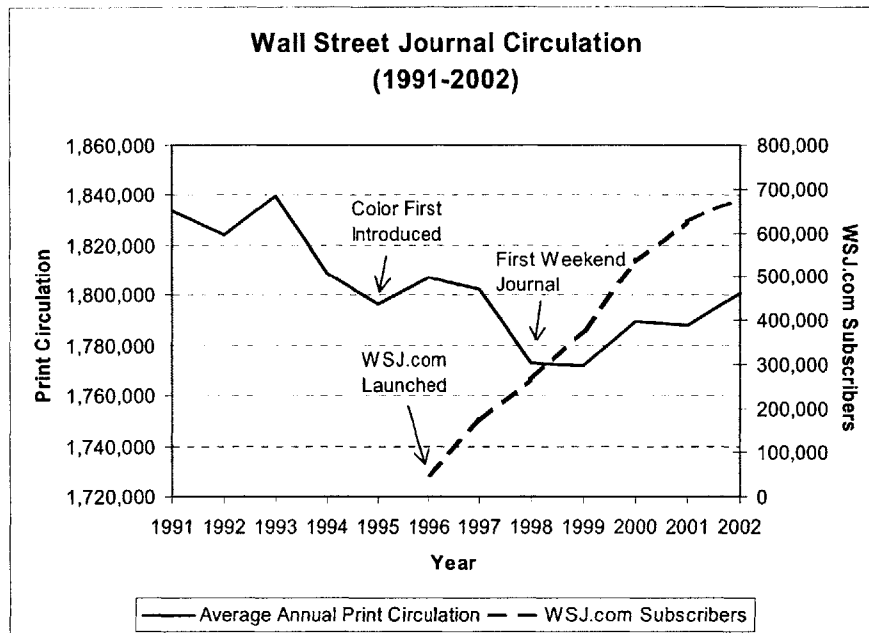
8. References

- Berry, Thomas D., and Keith M. Howe, 1994, "Public Information Arrival," *Journal of Finance*, Vol. 49, No. 4., pp. 1331-1346.
- Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 1998, "Investor Psychology and Security Market under- and Overreactions," *Journal of Finance*, Vol. 53, No. 6., pp. 1839-1885.
- Engle, Robert F., and Victor K. Ng, 1993, "Measuring and Testing the Impact of News on Volatility," *Journal of Finance*, Vol. 48, No. 5., pp. 1749-1778
- Hirshleifer, David, and Tyler Shumway, 2003, "Good day sunshine: Stock returns and the weather," *Journal of Finance*, Vol. 58, No. 3., pp. 1009-32.
- Hong, Harrison, Terence Lim, and Jeremy C. Stein, 2000, "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies," *Journal of Finance*, Vol. 55, No. 1., pp. 265-295.
- Lee, Charles M. C., Andrei Schleifer, and Richard H. Thaler, 1991, "Investor Sentiment and the Closed-End Fund Puzzle," *Journal of Finance*, Vol. 46, No. 1., pp. 77-109.
- Liu, Hugo, Henry Lieberman, and Ted Selker, 2003, "A Model of Textual Affect Sensing using Real-World Knowledge," in *Proc. Seventh International Conference on Intelligent User Interfaces (IUI 2003)*, pp. 125-132. Miami, Florida.
- Lo, Andrew, John Campbell, As. Craig MacKinlay, 1997, *The Econometrics of Financial Markets*, 1997, Princeton University Press. Princeton, NJ.
- Mitchell, Mark L., and J. Harold Mulherin, 1994, "The Impact of Public Information on the Stock Market," *Journal of Finance*, Vol. 49, No. 3, Papers and Proceedings Fifty-Fourth Annual Meeting of the American Finance Association, Boston, Massachusetts, January 3-5, 1994. pp. 923-950.
- Niederhoffer, Victor, 1971, "The Analysis of World Events and Stock Prices," *Journal of Business*, Vol. 44, No. 2., pp. 193-219.
- Pearce, Douglas K., and V. Vance Roley, 1985, "Stock Prices and Economic News," *Journal of Business*, Vol. 58, No. 1., pp. 49-67.
- Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates, 1966, "The General Inquirer: A Computer Approach to Content Analysis," *The MIT Press*.

Appendix A: The Wall Street Journal

Figure A.1 below shows the annual average print circulation of The Wall Street Journal and the annual subscription of WSJ.com for the past 12 years.⁷

Figure A.1: Wall Street Journal Circulation (1991-2002)



First issue of The Wall Street Journal published on July 8, 1889.⁸ Color was first introduced on October 22, 1995 in the form of color advertisement.⁹ Color was added to the rest of the newspaper on April 9, 2002. Also starting April 9, 2002, the Journal announced an increase in its page count from 80 to 96 with 24 color-capable pages, up from 8 previously. On the same date, the Journal also introduced a new section called “Personal Journal,” published every Tuesday, Wednesday, and Thursday.¹⁰ The electronic subscription WSJ.com started on April 29, 1996 and currently has 686,000 paid online subscribers.¹¹ The first weekend journal published on March 20, 1998.⁴

⁷ <http://www.dowjones.com/avgcirc.htm>, Dow Jones Annual Reports

⁸ <http://www.dowjones.com/factsht.htm>

⁹ <http://www.naa.org/technews/tn960506/p19color.html>

¹⁰ <http://www.bizjournals.com/tampabay/stories/2001/12/10/daily10.html>

¹¹ <http://www.ojr.org/ojr/glaser/1068601595.php>

Appendix B: Emotional Category Creation

Before proceeding with the methodology of emotional category creation, let's first define the input and output of this overall method. The input to the process is a list of words and a list of documents for each day. The list of words contains all words we'd like to examine as possible significant words. The list of documents is the training corpus we will use for computing various input word frequency statistics to determine word significance. It is a good idea to have a fairly large number of documents for statistical reasons that will soon become clear. The output of the method is a list of significant words on each day.

The process of identifying significant words on each day can be broken down into two major steps: noise filter and significance test. In the noise filter stage, we want to eliminate words that exhibit unpleasant qualities, such as occurring in high frequencies in a large number of documents. In the significance test stage, we would like to examine the non-noise words on each day and determine whether any are occurring more often than usual. We now proceed to describe the two stages more formally.

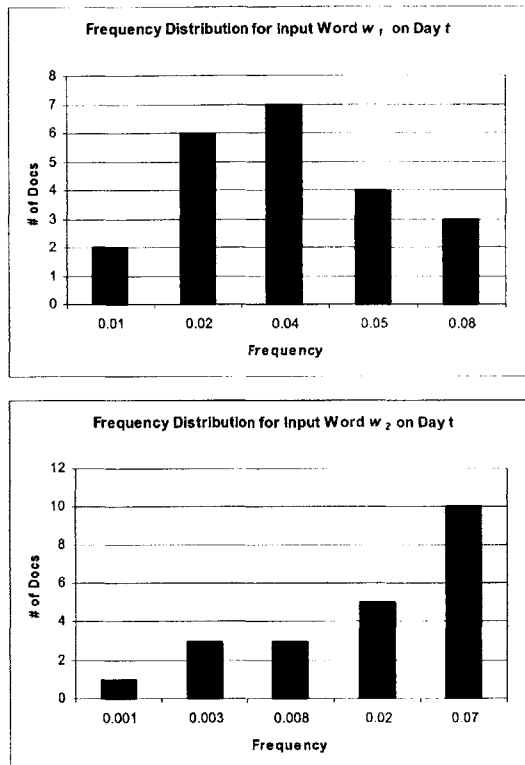
B.1. Noise Filtering

In order to determine which words among the input words list exhibit unpleasant behavior, we use the distribution of word frequencies. The frequency of word w_i in document j , $F_{w_i, j}$, is computed as the number of occurrences or count of w_i in document j , $Count(w_i, j)$, divided by the total number of words in document j .

$$F_{w_i, j} = \frac{Count(w_i, j)}{\sum_i Count(w_i, j)}$$

For each day, we construct for all input words found on that day a distribution that has on the horizontal axis the frequencies of that word and on the vertical axis the number of documents that contain the respective frequency of the word. Because it is very rare for two or more documents on any given day to have the same exact frequency for a word, we round each frequency number to the nearest non-zero digit after the decimal. This has the effect of grouping word frequencies into "bins" which in turn produces the frequency distribution. Two example distributions for input words w_1 and w_2 on a particular day t are shown in Figure B.1.

Figure B.1: Frequency distributions for input words w_1 and w_2 on day t



We then compute the kurtosis and skewness of each word frequency distribution and eliminate on each day words whose distribution fall outside a certain predefined kurtosis or skewness range. By doing so, we can eliminate input words that occur in high frequency in a large number of documents, as their distributions would have a high skewness and kurtosis. In the example distributions shown in Figure 1, word w_2 may be eliminated. We can also eliminate input words that occur in only one or two documents because its distribution will have a high kurtosis. Criteria other than a simple range test of kurtosis and skewness, such as nonlinear functions of those statistics, may be used to determine which words are noise words.

B.2. Significance Test

The noise filtering stage has produced a list of non-noise input words for each day from which we select any significant words using the significance test method. The general idea of the significance test is to pick out words that appear more often on a given day than average by a certain threshold.

The first set of statistics computed is the daily average frequency of each non-noise input word found on that day. The average frequency on a day t for a particular input word w_i , $\bar{F}_{w_i,t}$, is computed from the word frequency distribution used by the noise filter stage

as the sum of all the products of the word frequency on day t , $F_{w_i,t}$, and the number of documents that contain w_i in that frequency $D_{F_{w_i,t}}$, divided by the total number of documents on day t :

$$\bar{F}_{w_i,t} = \frac{\sum_F F_{w_i,t} \cdot D_{F_{w_i,t}}}{\sum_F D_{F_{w_i,t}}}$$

The next set of statistics to be computed is the average daily frequency and standard deviation of all input words (not just the non-noise input words) across all days and all documents. These average daily frequencies and standard deviation statistics form the benchmark that input words on each day will be measured against to determine their significance. We now express these benchmark statistics formally. Let d be the total number of days on which an input word w_i was found. The average daily frequency for word w_i , \bar{F}_{w_i} , is simply the sum of all daily frequencies, $\bar{F}_{w_i,t}$, divided by d . The frequency standard deviation S_{w_i} is computed using the same set of daily frequencies $\bar{F}_{w_i,t}$.

$$\bar{F}_{w_i} = \frac{\sum_t \bar{F}_{w_i,t}}{d}$$

The actual significance test is simply that if a word's average frequency on a particular day $\bar{F}_{w_i,t}$ is greater than or equal to its average daily frequency \bar{F}_{w_i} by a constant k times the frequency standard deviation S_{w_i} , then it is significant on that day:

$$\bar{F}_{w_i,t} \geq \bar{F}_{w_i} + k \cdot S_{w_i}$$

If an input word appears on only one day, it is automatically not considered significant because there is no benchmark daily average frequency to compare the word average frequency to, so we don't want to assume false significance.