

# ULRR

## Using financial event phrases and keywords to classify form 8-K disclosures by likely share price response

Item Type	Thesis
Authors	Slattery, Darina M.
Download date	2026-03-13 13:40:36
Item License	<a href="https://creativecommons.org/licenses/by-nc-sa/1.0/">https://creativecommons.org/licenses/by-nc-sa/1.0/</a>
Link to Item	<a href="https://hdl.handle.net/10344/6597">https://hdl.handle.net/10344/6597</a>



UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

Using Financial Event Phrases and Keywords to  
Classify Form 8-K Disclosures by Likely Share  
Price Response

*by*

Darina M. Slattery

Submitted in Fulfilment of the Requirements for the

Degree of Doctor of Philosophy

in

Computer Science

at

University of Limerick

Supervisor:

Dr. Richard F. E. Sutcliffe

Submitted to the University of Limerick, November 2012

# Abstract

## Using Financial Event Phrases and Keywords to Classify Form 8-K Disclosures by Likely Share Price Response

Darina M. Slattery

It is generally agreed that there are three different types of financial information: information in past stock prices, information that is available to all the public, and information that is both available to the public and available privately to insiders (Fama 1970; Haugen 1990; Hellstrom and Holmstrom 1998; Elton et al 2003). There is considerable debate about the possible impact that different kinds of information can have on the value of financial instruments. On the one hand, the efficient markets hypothesis (EMH) states that the price of a financial instrument properly reflects all available information immediately (Fama 1970). If security prices respond to all available information quickly, then the market is deemed efficient and no excess profits or returns can be made. On the other hand, fundamental and technical analysts argue that the market is inefficient because information disseminates slowly through the market and prices under- or over-react to the information (Haugen 1990).

A number of different data sources, features, goals, and methods have been used to automatically analyse content in financial documents. However, there has been very little research undertaken in the area of *automatic event phrase recognition and classification* of online disclosures. Our research study focuses on content contained in Form 8-K disclosures filed on EDGAR, a system maintained by the Securities and Exchange Commission (SEC). In our research study, we developed a prototype automatic financial event phrase (FEP) recogniser and we automatically classified a small sample of 8-Ks by likely share price response, using the automatically recognised FEPs and hand-chosen keywords as features. In our comparative classification experiments, we used the C4.5 suite of programs and the SVM-Light support vector machine program. Our datasets comprised 8-Ks filed by 50 randomly-chosen S&P 500 companies from 1997 to 2000 and 2005 to 2008.

Our research experiments yielded some interesting findings. In an experiment on the 2005 to 2008 dataset comprising 280 8-Ks, C4.5 was able to correctly classify 63.2% of the ‘ups’<sup>1</sup> (as against 58.2% at chance), when using FEPs and keywords. We also found that C4.5 appears to be better at identifying patterns in the training cases than SVM-Light, regardless of whether they were ‘ups’ or ‘downs’. When we compared the results from our FEP experiments with the results from two baseline approaches—n-gram classification and Naïve Bayes bag-of-words classification—we found that C4.5 using FEPs and keywords yielded marginally higher overall classification accuracy than C4.5 using n-grams or Naïve Bayes bag-of-words. A detailed description of the classification experiments is provided in the thesis, along with a discussion of the strengths and limitations of the research study. Recommendations for future work include further refinement of the FEPs and keywords, classification of larger datasets, and incorporation of additional classification variables beyond financial event phrases and hand-chosen keywords.

---

<sup>1</sup> By ‘ups’/ ‘downs’ we mean 8-Ks that had an increase/ decrease in share price around the filing date.

# Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted for any other academic award.

---

Signature of Author

---

Date

# Acknowledgements

I would first like to thank my supervisor, Dr. Richard F. E. Sutcliffe for supervising my project. His on-going interest in my progress kept me focused, even when other duties—including a full-time lecturing job and motherhood—were constantly vying for my attention. It has been a long and oftentimes a difficult project, but thankfully we have reached the end.

I would like to acknowledge Prof. Eamonn J. Walsh, University College Dublin, for providing guidance in the early stages and for providing an early dataset. I would also like to acknowledge Ms. Annette McElligott, Head of the Computer Science and Information Systems Department (CSIS), for facilitating my completion.

Special thanks to Dr. Kieran J. White, formerly of the CSIS department, for all his technical advice over the years. Thanks also to Dr. Michael P. O'Brien, formerly of CSIS and now a colleague in the School of LLCC, for the encouragement (and cups of tea) that helped me greatly through the final stages of the write-up. I would also like to acknowledge the help of Dr. Ailish Hannigan, Department of Mathematics and Statistics, for her advice during statistical consultations.

Lastly, but certainly not least, I would especially like to thank my family. Special thanks to my parents, Brian and Rose, for facilitating my education from day one and for being wonderful parents—this thesis is dedicated to them. Special thanks also to my brother Darwin who provided plenty of technical advice over the years, but more importantly, for providing much needed encouragement throughout. Special thanks to my husband Paulie for supporting me throughout the entire process. Finally, special thanks to our beautiful baby Holly, the light of our lives.

# Table of Contents

<b>ABSTRACT.....</b>	<b>II</b>
<b>DECLARATION .....</b>	<b>III</b>
<b>ACKNOWLEDGEMENTS.....</b>	<b>IV</b>
<b>LIST OF FIGURES.....</b>	<b>XI</b>
<b>LIST OF TABLES.....</b>	<b>XII</b>
<b>GLOSSARY OF FINANCIAL TERMS .....</b>	<b>XVI</b>
<b>CHAPTER 1: INTRODUCTION .....</b>	<b>1</b>
<b>1.1 Outline .....</b>	<b>1</b>
<b>1.2 Motivation .....</b>	<b>1</b>
<b>1.3 Research Objectives .....</b>	<b>2</b>
<b>1.4 Guide to Other Chapters .....</b>	<b>4</b>
<b>CHAPTER 2: EFFICIENT MARKETS, THE LANGUAGE OF NEWS, AND MARKET REACTIONS.....</b>	<b>6</b>
<b>2.1 Outline .....</b>	<b>6</b>
<b>2.2 Investment Analysis or the Efficient Markets Hypothesis? .....</b>	<b>8</b>
<b>2.3 The Language of Financial Reports and News.....</b>	<b>14</b>
<b>2.4 Information, Events and Market Reactions .....</b>	<b>29</b>

2.5 Summary .....	39
<b>CHAPTER 3: AUTOMATIC ANALYSIS AND CLASSIFICATION OF FINANCIAL DOCUMENTS .....</b>	<b>41</b>
3.1 Outline .....	41
3.2 Automatic Text Classification and Categorisation .....	42
3.3 Single Words.....	50
3.4 Keyword Records and Phrases.....	85
3.5 Financial Ratios and Variables .....	98
3.6 Summary .....	103
<b>CHAPTER 4: EVENTS IN FORM 8-K DISCLOSURES .....</b>	<b>108</b>
4.1 Outline .....	108
4.2 Background to Corporate Reporting on the EDGAR System .....	108
4.3 Datasets .....	116
4.3.1 The 7372 Dataset.....	116
4.3.2 The First S&P500 Dataset .....	119
4.3.3 The Second S&P500 Dataset.....	123
4.4 The Causal Effect of 8-K Disclosures in Various Windows .....	128
4.5 Identification of Financial Event Phrases (FEPs) .....	134
4.6 Summary .....	138
<b>CHAPTER 5: AUTOMATIC ANALYSIS OF FINANCIAL EVENTS IN 8-K DISCLOSURES .....</b>	<b>139</b>
5.1 Outline .....	139
5.2 Recognition of FEPs in Form 8-Ks .....	139
5.2.1 Development of the FEP Recogniser .....	140
5.2.2 Issues with the FEP Recogniser.....	141

<b>5.3 Automatic Pattern Analysis of FEPs</b> .....	<b>142</b>
5.3.1 Pattern Analysis of FEPs in the First S&P 500 Dataset .....	142
5.3.2 Pattern Analysis of FEPs in the Second S&P 500 Dataset.....	145
5.3.3 Discussion of FEP Patterns in Both Datasets .....	147
<b>5.4 Summary</b> .....	<b>157</b>
<b>CHAPTER 6: AUTOMATIC CLASSIFICATION OF 8-K DISCLOSURES USING VARIOUS DOCUMENT CONTENT FEATURES</b> .....	<b>158</b>
<b>6.1 Outline</b> .....	<b>158</b>
<b>6.2 Decision Tree Classification using Financial Event Phrases and Keywords</b> .....	<b>159</b>
6.2.1 Background to Decision Trees and C4.5.....	159
6.2.2 Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5) .....	163
6.2.3 Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5) .....	166
6.2.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5) ..	169
6.2.5 Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5) .....	172
6.2.6 Analysis and Discussion: C4.5 Results .....	175
<b>6.3 Support Vector Machine Classification using Financial Event Phrases and Keywords</b> .....	<b>180</b>
6.3.1 Background to Support Vector Machines and SVM-Light.....	180
6.3.2 Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light) .....	183
6.3.3 Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light) .....	184
6.3.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM- Light) .....	185
6.3.5 Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light) .....	187
6.3.6 Analysis and Discussion: SVM-Light Results .....	188
<b>6.4 Decision Tree Classification using N-Grams</b> .....	<b>192</b>
6.4.1 Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5) .....	192
6.4.2 Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5) .....	195
6.4.3 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5) ..	197
6.4.4 Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5) .....	199
6.4.5 Analysis and Discussion: C4.5 Results .....	202
<b>6.5 Support Vector Machine Classification using N-Grams</b> .....	<b>204</b>
6.5.1 Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light) .....	204
6.5.2 Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light) .....	205

6.5.3 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM-Light) .....	206
6.5.4 Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light).....	207
6.5.5 Analysis and Discussion: SVM-Light Results .....	208
<b>6.6 Naive Bayes Classification using a Bag-of-Words Approach .....</b>	<b>211</b>
6.6.1 Background to Naïve Bayes .....	211
6.6.2 Classification Results for the 1997-2000 Dataset: Subset 1 (Naïve Bayes).....	212
6.6.3 Classification Results for the 2005-2008 Dataset: Subset 2 (Naïve Bayes).....	213
6.6.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (Naïve Bayes) .....	214
6.6.5 Classification Results for the 2005-2008 Dataset: Subset 3 (Naïve Bayes).....	214
6.6.6 Analysis and Discussion: Naïve Bayes Results.....	215
<b>6.7 Summary .....</b>	<b>216</b>
 <b>CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS.....</b>	 <b>219</b>
<b>7.1 Outline .....</b>	<b>219</b>
<b>7.2 Summary of Key Findings.....</b>	<b>219</b>
<b>7.3 Strengths and Limitations of our Approach .....</b>	<b>223</b>
<b>7.4 Further Research .....</b>	<b>225</b>
<b>7.5 Summary .....</b>	<b>226</b>
 <b>REFERENCES .....</b>	 <b>227</b>
 <b>APPENDIX 1: HISTOGRAM FOR <math>T \pm 1</math> DAYS (1997-2000) .....</b>	 <b>243</b>
<b>APPENDIX 2: BOXPLOT FOR <math>T \pm 1</math> DAYS (1997-2000).....</b>	<b>244</b>
<b>APPENDIX 3: HISTOGRAM FOR <math>T \pm 5</math> DAYS (1997-2000) .....</b>	<b>245</b>
<b>APPENDIX 4: BOXPLOT FOR <math>T \pm 5</math> DAYS (1997-2000).....</b>	<b>246</b>
<b>APPENDIX 5: HISTOGRAM FOR <math>T \pm 100</math> DAYS (1997-2000) .....</b>	<b>247</b>

<b>APPENDIX 6: BOXPLOT FOR T±100 DAYS (1997-2000)</b> .....	<b>248</b>
<b>APPENDIX 7: HISTOGRAM FOR T±1 DAYS (2005-2008)</b> .....	<b>249</b>
<b>APPENDIX 8: HISTOGRAM FOR T±1 DAYS (2005-2008)</b> .....	<b>250</b>
<b>APPENDIX 9: HISTOGRAM FOR T±5 DAYS (2005-2008)</b> .....	<b>251</b>
<b>APPENDIX 10: BOXPLOT FOR T±5 DAYS (2005-2008)</b> .....	<b>252</b>
<b>APPENDIX 11: HISTOGRAM FOR T±100 DAYS (2005-2008)</b> .....	<b>253</b>
<b>APPENDIX 12: BOXPLOT FOR T±100 DAYS (2005-2008)</b> .....	<b>254</b>
<b>APPENDIX 13: TIME SERIES FOR HALLIBURTON CO</b> .....	<b>255</b>
<b>APPENDIX 14: TIME SERIES FOR MORGAN STANLEY DEAN WITTER &amp; CO.</b> .....	<b>256</b>
<b>APPENDIX 15: TIME SERIES FOR NIKE INC</b> .....	<b>257</b>
<b>APPENDIX 16: TIME SERIES FOR RED HAT INC</b> .....	<b>258</b>
<b>APPENDIX 17: TIME SERIES FOR LINCOLN NATIONAL CORP</b> .....	<b>259</b>
<b>APPENDIX 18: TIME SERIES FOR O'REILLY AUTOMOTIVE INC</b> .....	<b>260</b>
<b>APPENDIX 19: TIME SERIES FOR DR HORTON INC</b> .....	<b>261</b>
<b>APPENDIX 20: SAMPLE PHRASES FOR EMPLOYMENT-RELATED FEPS</b> .....	<b>262</b>
<b>APPENDIX 21: LIST OF FEP TYPES, NAMED ENTITIES, AND TYPES OF FINANCIAL OBJECT</b> .....	<b>264</b>
<b>APPENDIX 22: LIST OF KEYWORDS</b> .....	<b>267</b>

**APPENDIX 23: DIFFERENT FEP TYPES RECOGNISED IN THE FIRST S&P 500 DATASET ..... 283**

**APPENDIX 24: DIFFERENT FEP TYPES RECOGNISED IN THE SECOND S&P 500 DATASET ..... 285**

**APPENDIX 25: DIFFERENT FEP TYPES APPEARING IN C4.5 DECISION TREES ..... 287**

# List of Figures

Figure 3.1: Recall-precision graph (Frakes and Baeza-Yates 1992 p.11).....	46
Figure 5.1: The FEP, NE, and TFO recognition process. ....	140
Figure 6.1: Sample content for the C4.5 .names file (keywords, FEPs, named entities, and types of financial object). ....	161
Figure 6.2: Sample content for the C4.5 .data training file (two cases).....	162
Figure 6.3: Sample content for the C4.5 .test test file (two cases).....	162
Figure 6.4: Sample content for the SVM-Light .data training file (two cases).....	182
Figure 6.5: Sample content for the SVM-Light .test test file (two cases).....	182

# List of Tables

Table 2.1: Data sources, document features examined, goals, and methods used in various studies that examined the language in annual and interim reports and news articles. ....	17
Table 2.2: Typical words used to signify different themes (Kohut and Segars 1992).	21
Table 3.1: Data sources, document features examined, goals, and methods used for the automatic analysis and classification of financial documents.....	49
Table 3.2: Breakdown for each textual representation (Schumaker and Chen 2006).	55
Table 3.3: Average standardized cumulative abnormal returns for news events, scaled by 100 for readability purposes (Antweiler and Frank 2006). ....	62
Table 3.4: Inverse document frequency of one positive and one negative substring in the basic word set (Kroha and Baeza-Yates 2004). ....	71
Table 3.5: Inverse document frequency of one positive and one negative substring with high probabilities (Kroha and Baeza-Yates 2004). ....	72
Table 3.6: Breakdown of news items for each trend set (Kroha et al 2006). ....	73
Table 3.7: Definition of hard and soft content of earnings news (Engelberg 2008)...	77
Table 3.8: Methods used to generate word lists (various studies). ....	85
Table 3.9: Results on test data without cross-validation and combining between one and seven sources (Cho et al 1999). ....	91
Table 3.10: Results on test data using five-fold cross validation and combining between one and seven sources (Cho et al 1999). ....	92
Table 3.11: Results of neural network experiments using different years, variables, training, and test sets (Lam 2004). ....	102
Table 3.12: Methods, evaluation metrics, and research considerations used for the automatic analysis and classification of financial documents.....	105
Table 3.13: Three rigorous studies determined by the range of evaluation metrics used and the robustness of their research methods. ....	107
Table 4.1: Initial listing of reportable event items in Form 8-Ks (SEC 2002).....	110
Table 4.2: Amended listing of reportable items in Form 8-Ks (SEC 2004a). ....	111
Table 4.3: Subsequent amended listing of reportable event items in Form 8-Ks (SEC 2004b; SEC 2005; SEC 2009a; SEC 2011). ....	112

Table 4.4: Characteristics of data used in the 7372 dataset (number of newlines, words, and size).....	119
Table 4.5: Initial random listing of our 50 S&P 500 companies. ....	120
Table 4.6: Characteristics of data used in the first S&P dataset (number of newlines, words, and size).....	123
Table 4.7: Issues and solutions for the revised listing of S&P companies.....	125
Table 4.8: Characteristics of data used in the second S &P dataset (number of newlines, words, and size).....	126
Table 4.9: Final random listing of S&P 500 companies. ....	127
Table 4.10: Range of share price changes for various windows (both datasets, available prices).....	128
Table 4.11: 90% of data lies within these ranges (both datasets, available prices). .	129
Table 4.12: Initial ontology from the manual analysis of disclosures. ....	134
Table 4.13: Final ontology of financial event phrases (FEPs) used in the grammar.	136
Table 5.1: Total, average, and range of FEPs recognised in all ‘downs’ and ‘ups’, including duplicates (first S&P 500 dataset).....	143
Table 5.2: Number of ‘downs’ and ‘ups’ which had 0 or more FEP types (first S&P 500 dataset). ....	143
Table 5.3: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (first S&P 500 dataset).....	144
Table 5.4: Number of ‘downs’ and ‘ups’ with unique FEPs in ‘downs’ and ‘ups’, no duplicates (first S&P 500 dataset).....	144
Table 5.5: Total, average, and range of FEPs recognised in all ‘downs’ and ‘ups’, including duplicates (second S&P 500 dataset). ....	145
Table 5.6: Number of ‘downs’ and ‘ups’ which had 0 or more FEP types (second S&P 500 dataset).....	146
Table 5.7: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (second S&P 500 dataset).....	146
Table 5.8: Number of ‘downs’ and ‘ups’ with unique FEPs in ‘downs’ and ‘ups’, no duplicates (second S&P 500 dataset). ....	147
Table 5.9: Number of words in ‘downs’ and ‘ups’ (both datasets).....	148
Table 5.10: Total, average, and range of FEPs recognised in all ‘downs’ and ‘ups’, including duplicates (both datasets).....	149

Table 5.11: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (both datasets).	151
Table 5.12: Occurrences of each FEP type in ‘downs’ and ‘ups’ (both datasets).	153
Table 5.13: Most frequently-occurring FEP types in ‘downs’ and ‘ups’, no duplicates (both datasets).	154
Table 5.14: FEP types that occurred together in ‘downs’ and ‘ups’ (both datasets).	155
Table 5.15: Most frequently-occurring FEP types in ‘downs’ and ‘ups’ (both datasets).	156
Table 6.1: Number of errors and tree sizes for the 1997-2000 dataset: subset 1 (C4.5).	164
Table 6.2: Confusion matrix for the 1997-2000 dataset: subset 1 (C4.5).	166
Table 6.3: Number of errors and tree sizes for the 2005-2008 dataset: subset 2 (C4.5).	167
Table 6.4: Confusion matrix for the 2005-2008 dataset: subset 2 (C4.5).	168
Table 6.5: Number of errors and tree sizes for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).	170
Table 6.6: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).	171
Table 6.7: Number of errors and tree sizes for all cases in the 2005-2008 dataset: subset 3 (C4.5).	173
Table 6.8: Confusion matrix for the 2005-2008 dataset: subset 3 (C4.5).	174
Table 6.9: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 dataset: subset 1 (SVM-Light).	184
Table 6.10: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 2 (SVM-Light).	185
Table 6.11: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (SVM-Light).	186
Table 6.12: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subsets 3(SVM-Light).	187
Table 6.13: Number of errors and tree sizes for the 1997-2000 dataset: subset 1 (C4.5).	193
Table 6.14: Confusion matrix for the 1997-2000 dataset: subset 1 (C4.5).	194
Table 6.15: Number of errors and tree sizes for the 2005-2008 dataset: subset 2	

(C4.5).....	196
Table 6.16: Confusion matrix for the 2005-2008 dataset: subset 2 (C4.5).....	197
Table 6.17: Number of errors and tree sizes for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).....	198
Table 6.18: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).....	199
Table 6.19: Number of errors and tree sizes for the 2005-2008 dataset: subset 3 (C4.5).....	200
Table 6.20: Confusion matrix for the 2005-2008 dataset: subset 3 (C4.5).....	201
Table 6.21: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 dataset: subset 1 (SVM-Light).....	205
Table 6.22: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 2 (SVM-Light).....	206
Table 6.23: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (SVM-Light).....	207
Table 6.24: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 3 (SVM-Light).....	208
Table 6.25: Accuracy levels of SVM-Light and n-grams compared to a arbitrary classification (all experiments).....	209
Table 6.26: Confusion matrix for the 1997-2000 dataset: subset 1 (Naïve Bayes) ..	213
Table 6.27: Confusion matrix for the 2005-2008 dataset: subset 2 (Naïve Bayes) ..	213
Table 6.28: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (Naïve Bayes).....	214
Table 6.29: Confusion matrix for the 2005-2008 dataset: subset 3 (Naïve Bayes) ..	215
Table 6.30: Overall accuracy of ups and downs together, for all four experiments (Naïve Bayes).....	215
Table 6.31: Overall accuracy of ups and downs separately, for all four experiments (Naïve Bayes).....	215
Table 6.32: Average classification accuracy for all four experiments using C4.5, SVM-Light, and Naïve Bayes, with various document content features (FEPs and keywords, n-grams, and a bag-of-words).....	218

# Glossary of Financial Terms

## **Abnormal return**

The difference between the expected (normal) return on an investment and the actual return. The normal return is due to market-wide influences.

## **Annret**

The percentage share price return around the announcement date. For a three-day window (days  $t \pm 1$ ), the annret is calculated as follows:  $((P_{t+1})/(P_{t-1}))-1$ , where P stands for the closing share price.

## **EDGAR**

The Electronic Data Gathering, Analysis, and Retrieval System. The EDGAR system is maintained by the Securities and Exchange Commission (SEC) and provides free public access to corporate information, including Form 8-K disclosures.

## **Excess return**

Return from an investment that exceeds some benchmark or index with a similar level of risk. Also known as the alpha.

## **Going-concern**

A company that has sufficient resources to enable it to continue to operate indefinitely and not go bankrupt.

## **Rate of return**

The amount of revenue (interest or dividend) that is generated by an investment, shown as a percentage of the original capital invested.

## **Round-trip**

This is when an investor takes a particular position and then closes it later. An example of a round trip is when an investor buys a futures contract and then sells it. Also called a round turn.

**Securities and Exchange Commission (SEC)**

The US federal agency created by the Securities Exchange Act of 1934 to enforce federal securities laws. The SEC is charged with promoting the disclosure of important market-related information, protecting investors against fraud, and maintaining fair dealing.

**Spread**

The difference between the bid and ask (offer) prices for an asset. The bid-ask spread is the difference between the highest price a buyer is willing to pay and the lowest price a seller is willing to offer.

**Standardised unexpected earnings (SUE)**

The difference between the median of the analysts' forecasts and the firm's actual earnings per share, scaled by a normalisation factor (Engelberg 2008).

**Switching portfolio/ strategy**

An investment strategy whereby investment funds are switched between bonds and shares, depending on whether the excess return is likely to increase or decrease.

**Transaction costs**

Any costs incurred when buying or selling securities. Costs may include brokers' commission fees and the bid-ask spread.

**Volatility**

A measure of the stability (or otherwise) of a security. Highly-volatile securities are considered high-risk as they may perform very well or very poorly.



# Chapter 1: Introduction

## 1.1 Outline

We begin this chapter by explaining the motivation for our research. We then describe our primary and secondary research objectives, outlining any underlying assumptions we made. Our introduction ends with a synopsis of the remaining chapters.

## 1.2 Motivation

Whilst it is the general consensus amongst investment analysts and investors that financial news impacts share prices, advocates of the efficient markets hypothesis would argue that publicly-available financial information cannot be used for prediction purposes. Even if short-term profits can be made, they would argue that no-one will be able to *consistently* outperform the market using fundamental information about a company (Malkiel 2007). Nonetheless, this latter debate has not lessened the enthusiasm of those wishing to find the perfect investment solution. In addition, many who believe that information has a predictive power differ in opinion regarding how long it takes for the market to digest news. Fama (1991) reported that, on average, stock prices “adjust within a day to event announcements” (p.28).

A significant amount of research to-date has focused on time series prediction (see, for example, Tay et al 2003 for a good review). However, with the massive increase in textual information that is now available online, more recent studies have focused on the language of financial reports and news. The goals of these studies have included narrative analysis, quantitative and qualitative analysis of content, writing style and tone analysis, thematic analysis, effectiveness analysis, performance and readability analysis, positive and negative word analysis, and market sentiment analysis (see Section 2.3 for a discussion of some relevant studies). In Chapter 3, we describe studies that used automatic methods for content analysis and classification of financial documents. These studies used various data sources, examined different features, had different goals, and used different methods.

Previous research undertaken by Slattery et al (2002) investigated Form 8-K disclosures and the potential for using five-word compound phrases to automatically predict share price reactions to those 8-Ks; this early research has since led us to explore the area of automatic *event phrase* recognition in 8-Ks, an area not previously explored in other research. We chose Form 8-Ks because they must be filed within a few days of certain material events and they are not as susceptible to noise as other online sources. 8-Ks also provide more content for analysis, unlike news headlines (see, for example, Thomas 2003) and message boards (e.g. Antweiler and Frank 2004). We decided to examine and recognise *event phrases* because the majority of other studies examined single words (e.g. Tetlock et al 2008) or other types of keyword records and phrases (e.g. Gillam et al 2002).

This thesis describes (1) the development of a prototype financial event phrase (FEP) recogniser that could be used by investors and analysts in their toolbox to automatically analyse 8-Ks and (2) the classification of 8-K disclosures by likely share price response (up or down) using those FEPs and other features as input. We also compare the results from the FEP experiments, with two baseline approaches—n-gram classification and Naïve Bayes bag-of-words classification.

### **1.3 Research Objectives**

The primary objective of this project was to investigate if the use of FEPs, and possibly other features, could be used to automatically classify Form 8-Ks by likely share price response (up or down). Before we could automatically classify 8-Ks, it was necessary to automatically recognise FEPs in those disclosures.

We did not aim to prove that the efficient markets hypothesis holds true or that FEPs are the sole cause of share price changes; rather, we wanted to investigate the *potential usefulness* of FEPs, and possibly other features, in the automatic classification of 8-Ks.

A number of assumptions underlie our research:

- The market is not fully efficient. Prices do not always fully reflect all available information immediately.
- There is some value in fundamental data filed in 8-Ks (i.e. there can be a correlation between 8-K content and share price changes around the filing dates).
- Prices adjust within a day (Fama 1991) to news filed in Form 8-K disclosures.
- All companies are equally affected by external factors (we do not control for other variables).
- Firms voluntarily disclose positive news, even when they are not legally obliged to do so (Skinner 1994).
- Positive or good news is easier to classify automatically than negative news, as the latter is often disguised in the midst of positive news. Even though positive news is often provided voluntarily (see previous assumption), we do not assume that this reduces the impact of positive news, as compared to negative news.
- Firms only disclose negative news when they are legally obliged to do so (Liu 2000).

In addition to our primary objective, we also wished to explore a number of secondary objectives or questions:

- Are 8-Ks with an increase in share price around the filing date (hereafter referred to as the ‘ups’) easier to classify than 8-Ks with a decrease in share price (the ‘downs’)?
- Are 8-Ks filed after the 2004 SEC rule changes easier or more difficult to classify than 8-Ks filed beforehand? In 2004, the SEC brought out new regulations specifying additional items that must be disclosed in 8-Ks (see Section 4.2 for a breakdown of Form 8-K items) and they also reduced the filing deadline from ten to four days.
- Using our prototype FEP recogniser, are more event types recognised post-2004?
- Do certain event types occur more frequently in the ‘ups’ or ‘downs’?

## 1.4 Guide to Other Chapters

Chapter 2—Efficient Markets, the Language of News, and Market Reactions—briefly introduces various topics related to financial analysis, including types of financial securities, types of markets, popular measures for determining the value of a financial instrument, and types of financial information. The chapter then provides a more detailed overview of the literature in key related areas: market efficiency theory and investment analysis, the language of financial reports and news, and the relationships between information, financial events, and market reactions.

Chapter 3—Automatic Analysis and Classification of Financial Documents—is the second literature review chapter. The chapter begins with an overview of automatic text classification and categorisation methods. It then looks specifically at studies that used automatic techniques to analyse the textual content of financial documents, as we also used automatic techniques for this purpose. As the data sources, features examined, goals, and methods employed varied in these studies, we first describe studies that used single words, and possibly other features, for the automatic content analysis of financial documents. We then discuss studies that used keyword records and phrases, but not single words, for analysis. Finally, we describe studies that used financial ratios and variables. We do not discuss literature on financial time series prediction, as that is beyond the scope of our research.

Chapter 4—Events in Form 8-K Disclosures—presents the rationale for choosing Form 8-K disclosures as our research data. We briefly examine the Securities and Exchange Commission’s legal requirements regarding the filing of corporate disclosures and then discuss the features of 8-Ks in detail. We then outline some of the features of the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system, which hosts 8-Ks. We describe the characteristics of our datasets in detail, before exploring the causal effect of 8-K disclosures in various windows. Finally, we explain how we identified the financial event phrases (FEPs) that were subsequently used to recognise events in 8-Ks.

Chapter 5—Automatic Analysis of Financial Events in 8-K Disclosures—describes how we developed our prototype FEP recogniser and outlines some issues

encountered when developing the recogniser. We then present an overview of the features of the recognised output in each of the two main datasets, using automatic pattern analysis techniques. We then discuss the features of both datasets together, with a view to identifying possible trends or patterns in the ‘ups’ and ‘downs’.

Chapter 6—Automatic Classification of 8-K Disclosures using Various Document Content Features—presents the results from various experiments which used different methods (decision trees, support vector machines, and Naïve Bayes), as well as different content features (FEPs and keywords, n-grams, and a bag-of-words) for the classification of 8-Ks by likely share price response. After we present the results, we analyse and discuss the findings from each set of experiments.

Chapter 7—Conclusions and Recommendations—presents a summary of our key findings. We outline the strengths and weaknesses of our approach and highlight some important contributions of our research. Finally, we recommend areas worthy of future research.

# Chapter 2: Efficient Markets, the Language of News, and Market Reactions

## 2.1 Outline

In the first section of this chapter, we will briefly introduce various topics related to financial analysis. These topics will include types of financial securities, types of markets, popular measures for determining the value of a financial instrument, and types of financial information. The remaining sections will provide a more detailed overview of the literature in key related areas. Section 2.2 will discuss market efficiency theory and investment analysis, Section 2.3 will describe various papers which examined the language of financial reports and news, and Section 2.4 will outline relationships between information, financial events, and market reactions. Finally, Section 2.5 will provide a summary of this chapter.

There are many types of financial securities, including corporate stocks or equities (e.g. preferred stock and common stock), government bonds (e.g. treasury bills and treasury notes), corporate securities (e.g. debenture bonds and convertible bonds), options, warrants, and forward and futures contracts (Haugen 1990; Elton et al 2003). Some financial securities can have a fixed income (e.g. government bonds) whereas others can have a variable income (e.g. company common stocks). Our research is concerned with corporate stocks.

There are two types of markets – primary markets and secondary markets (Elton et al 2003; Haugen 1990; O'Loughlin and O'Brien 2006). Securities are initially sold on primary markets, such as the Federal Reserve in the US, either by the government or by a company that wishes to fund capital investment, and then later bought and sold by investors on secondary markets such as the New York Stock Exchange (NYSE), the London Stock Exchange (LSE), or the Irish Stock Exchange (ISE). The NYSE is the largest secondary market exchange in the US. To be traded on an exchange, a company must be listed on that exchange. Listing depends on various factors including the size of the company and former profits (Elton et al 2003; O'Loughlin and O'Brien 2006). Unlisted stocks, on the other hand, have to be traded in an over-

the-counter market (OTC). The NYSE has an automated system that lists all the OTC securities; this system is called the National Association of Security Dealers Automatic Quote System (NASDAQ). Originally this system was not a trading system—only an electronic information system—but it has since become the second largest stock market exchange in the US.

The performance of investments can be compared with the overall market performance using stock market indices, such as the Standard and Pools (S&P) 500 index in the US, the ISEQ Overall index in Ireland, and the FTSE 100 in the UK (O'Loughlin and O'Brien 2006). Our research uses the S&P 500 index to avoid industry bias, as this index comprises 500 leading companies in the US economy, across a broad spectrum of industries.

Typically, investors and analysts use statistical measures to determine the value of a particular financial investment. The most common measures include volatility which measures the standard deviation of the closing price from its average value in the past few days/hours/minutes; the moving average; and the return value which measures the difference between the value of the instrument at time  $t-1$  and at time  $t$  (Ahmad et al 2003). Another measure that has been used in more recent times is the frequency of occurrence of sentiment indicators (e.g. see Ahmad et al 2003 in Section 2.3). Our research study uses a version of the return value.

It is generally agreed that there are three different types of financial information: information in past stock prices, information that is available to all the public, and information that is both available to the public and available privately to insiders (Fama 1970; Haugen 1990; Hellstrom and Holmstrom 1998; Elton et al 2003). There is considerable debate about the possible impact that different kinds of information can have on the value of financial instruments. For example, the theory of market efficiency, or the efficient markets hypothesis (EMH), states that the price of a financial instrument properly reflects all available information immediately (Fama 1970). If security prices respond to all available information quickly, then the market is deemed efficient and no excess profits or returns can be made. Even though the EMH has many advocates, investment analysis is still a thriving and, oftentimes, a

profitable industry. The two main kinds of investment analyst – fundamental and technical – would argue that the market is inefficient because information disseminates slowly through the market and prices under- or over- react to the information (Haugen 1990). In Section 2.2, we will further explore the two sides to the debate, firstly by discussing the two kinds of investment analysis and then by discussing the EMH in more detail.

## **2.2 Investment Analysis or the Efficient Markets Hypothesis?**

Fundamental analysis involves evaluating the value of a financial instrument using quantitative and qualitative content derived from company financial statements, news stories, analysts' reports, and message forums. Relevant fundamental content can include ratios and variables such as the earnings per share (EPS), net profit, and trading volume (e.g. see Reinganum 1988; Ou 1990; Francis et al 2002; Lam 2004; Thomsett 2007) as well as announcements relating to company policies and possible structural changes (see Section 2.3 for a review of some studies which examined the language of financial reports and news as well as Chapter 3 for a detailed review of studies that involved automatic content analysis of financial documents). Frequently used online news sources include Barrons<sup>2</sup>, Bloomberg; Raging Bull, Reuters, Silicon Investor, the Motley Fool, the Financial Times, the Wall Street Journal, and Yahoo! Finance. Other sources of quantitative and qualitative information include online financial analysis tools, such as Datastream<sup>3</sup>, Dow Jones News Analytics, MarketScope Advisor, and VectorVest, and online databases, such as the Sagemworks Database Platform<sup>4</sup> and the Securities and Exchange Commission's EDGAR system. EDGAR hosts crucial corporate filings such as the annual Form 10-K and interim Form 8-K reports. Our research focuses solely on qualitative content found in Form 8-K reports for reasons outlined in Chapter 4.

---

<sup>2</sup><http://online.barrons.com/home-page>; <http://www.bloomberg.com/>; <http://www.ragingbull.com>; <http://www.reuters.com>; <http://siliconinvestor.advfn.com>; <http://www.fool.com>; <http://www.ft.com/home/europe>; <http://www.wsj.com>; <http://finance.yahoo.com/>;

<sup>3</sup><http://online.thomsonreuters.com/datastream/>; <http://www.dowjones.com/product-news-analytics.asp>; <http://www.marketscope.com/>; <http://www.vectorvest.com/products.htm>

<sup>4</sup><https://www.sageworksdatabase.com/productoverview.aspx>; <http://www.sec.gov/edgar.shtml>

Technical analysis, on the other hand, involves evaluating time series patterns and trends relating to previous prices of the financial instrument and the volume of trading, with a view to predicting future prices and volumes. Time series is beyond the scope of our research and it has already been researched extensively (e.g. see White 1988, Andersen and Bollerslev 1998a, Andersen and Bollerslev 1998b, and Lam 2004 for some relevant studies).

There are three forms of the efficient markets hypothesis (EMH): the weak form, the semi-strong form, and the strong form (Fama 1970; Haugen 1990; Hellstrom and Holmstrom 1998; Elton et al 2003; Malkiel 2007). Each form of the EMH deals with a different type of financial information, as outlined in Section 2.1. Whilst our research is only concerned with the second type (i.e. qualitative information that is available to all the public in Form 8-K reports) and assumes that the market is not fully efficient, we must still consider the three forms of the EMH and the various debates that surround them, because they may help explain our findings.

Analysts who believe the EMH uses statistical tests to show that the relevant information has no predictive power. Tests of the weak form assume that past prices cannot be used to predict future prices; in other words, they assume that technical analysis is useless because prices move in a random walk fashion (Fama 1965). Malkiel (2007), a proponent of the random walk theory<sup>5</sup>, elaborates further by saying "the history of stock price movements contains no useful information that will enable an investor consistently to outperform a buy-and-hold strategy in managing a portfolio" (*ibid*, p.132). He argues that the transaction costs<sup>6</sup> associated with trying to take advantage of any minor trends that might be identified, will be greater than any profits that could be made.

---

<sup>5</sup> Lo and MacKinlay (1999) said that while the EMH and random walk theory are often viewed as closely related, unpredictable prices do not automatically arise from an efficient market with rational investors; likewise, predictable prices do not automatically arise from an inefficient market with irrational investors.

<sup>6</sup> Transaction costs, such as broker commissions, arise when an investor buys or sells a financial instrument.

Tests of the semi-strong form of the hypothesis assume that stock prices already reflect not only past time series, but also information in publicly available company reports and information in government (macro) announcements (Fama 1970; Malkiel 2007). Numerous company-related characteristics have been linked to excess returns, including the "size effect" (Banz 1981), book-to-market value (Fama and French 1992), and earnings/price ratio (E/P ratio) (Basu 1977). However, if one assumes that the semi-strong form of the hypothesis is true, technical analysis of previous patterns will not prove useful; likewise, a fundamental analysis of company statements and other sources of publicly available news about the company will not yield excess returns because both buyers and sellers have access to the same information and will reassess the value of the security accordingly. Occasionally buyers may assess the impact of the news quicker than others and therefore they might be able to buy at a low price; likewise, others may take longer to digest the information and will probably pay a higher price (Elton et al 2003). Either way, advocates of the EMH believe that no one will be able to *consistently* outperform the market using fundamental information (Malkiel 2007). Whilst our research assumes that the market is inefficient to some extent, we do not aim to prove or disprove the EMH; rather, we are attempting (1) to identify financial event phrases in 8-Ks and (2) to use them in classification experiments to predict the *likely* share price response.

Tests of the strong form of the hypothesis assume that prices take all available information into account, including "inside" or private information (Fama 1970; Malkiel 2007). Obviously, only certain people can have access to this latter information and they tend to act on that information as soon as they receive it, with a view to making excess profits, even though this is against the law. Prices quickly reflect this information, so advocates consider this the most efficient form of the model. In reality, no security analyst can consistently rely on receiving inside information and even if an analyst does earn excess returns, one cannot be sure if these were generated due to the inside information, chance, or simply superior use of publicly available information (Fama 1970; Elton et al 2003; Malkiel 2007).

In his 1970 study of theory and empirical evidence, Fama found that the weak and semi-strong forms were reasonably sound. However, he suggested that the strong

form of the hypothesis should probably be used only as a "benchmark against which deviations from market efficiency... can be judged" (Fama 1970, p.415). Whilst zero transaction costs are highly unlikely in reality, and many argue that it is a fundamental flaw in Fama's theory, he found it useful to be able to ignore transaction costs when comparing the different forms of the model. Instead, he became more concerned with the meaning of "properly reflects" in the standard definition of the hypothesis and said that it could really only be tested in conjunction with some other equilibrium pricing model that defines the meaning of the word "properly". In a subsequent study over twenty years later, his focus changed even more. Instead, his goal was to show that market efficiency theory can be useful when trying to describe the behaviour of returns—in other words, he was less interested in proving how precise the model was (Fama 1991). Whilst our research assumes that prices do not always fully reflect all publicly available information, the EMH might go some way to explain the various findings from our experiments (see Chapter 6).

In a subsequent study, Fama (1991) modified some of the EMH classifications. Tests of the weak form hypothesis became known as 'tests of return predictability' and they now cover a wider area. Tests of the semi-strong form hypothesis are now referred to as 'event studies' or 'studies of announcements' and tests of the strong form are now referred to as 'tests for private information'. Concerning return predictability, most of the controversy about market efficiency centres on this form of the model. Fama found that work on short-horizon stock returns was still similar to earlier work but more precise than newer research on long-horizon stock returns (Fama 1991). With regard to event studies, Fama summarised some of the main event study findings in his 1991 paper (see Section 2.4) and concluded that event studies "can give a clear picture of the speed of adjustment of prices to information"<sup>7</sup>. Our research examines events in Form 8-K disclosures and explores closing price market reactions within a three-day window ( $t-1$  to  $t+1$ , where  $t$  is the 8-K filing date). With regard to tests for private information, he concluded that it is difficult to measure abnormal returns over long periods and to decipher what exactly caused those abnormal returns. He also said that there has been a marked increase in the number of professional fund

---

<sup>7</sup> See also MacKinlay (1997) for a review of event studies in economics and finance.

managers who follow passive strategies, which suggests that private information is not as abundant as previously thought.

Malkiel (2007) proposed five factors that he believed help explain why security analysts (both fundamental and technical) encounter difficulties when trying to consistently predict the future (*ibid.*, pp.155-156):

- The impact of random events.
- Creative accounting procedures, which can lead to misleading reports.
- Incompetent financial analysts.
- The loss of the best analysts to other investment roles.
- Conflicts of interest between analysts and their employers (e.g. banks).

The impact of random events and creative accounting procedures are possibly two reasons for some of our findings (see Chapter 6 for a detailed discussion).

Even though the EMH has been widely debated in academic circles and it has greatly influenced real-world practice (Fama 1991), it is still in the interests of investment analysts to convince others that the market is inefficient. Many like to propose that they can identify and interpret useful information in a more efficient manner than others, using technical and/or fundamental analysis, and that they can act quickly on it to generate excess returns (Hellstrom and Holmstrom 1998). Lo and MacKinlay (1999) argued that markets are predictable to some extent but that the predictability comes with its own costs; there is always a trade-off between risk and return. They argued that the EMH is an insufficient hypothesis that really needs to take other hypotheses into account; possible hypotheses could relate to investor preferences, the structure of the information, and the business conditions.

Brunnermeier (1998) examined how traders with inside or private information signals attempt to manipulate the price with a view to enhancing the informational advantage when the information is later released to the public. If other traders conduct a technical analysis after the announcement is made, they will probably make incorrect inferences about the value of the announcement as the early-informed trader has already manipulated the price by buying on rumours and selling on news. The early-informed trader can better interpret the past price as he/she knows the *extent* to which

the price already reflects the new public information. These traders trade for both manipulative and speculative reasons. Prices react to all this buying and selling, not just to the original information. Bagnoli et al (1999) compared earnings forecast "whispers" from various online sources with the forecasts from First Call analysts and found that whispers tended to be more accurate proxies than First Call forecasts, when they examined the returns that could possibly have been yielded (an average size-adjusted cumulative return of 2.90% vs. 1.03%). They also found that the whispers tended to overestimate earnings whereas the First Call analysts tended to underestimate earnings. Whilst the First Call estimations tend to be released much earlier than the whispers, they found that the whispers tended to have a slight timing advantage; this advantage was weakened, however, when they used First Call estimations that were released closer to the earnings announcement than the whispers. In addition, they found that information tends to flow from whispers into the price, not the other way around, which suggests that whispers from various online sources are timely sources of information.

Some critics of the EMH proclaim that the theory states that stock prices are insensitive to changes in fundamental information; Malkiel (2007) argues to the contrary by saying that the EMH means that "the market is so efficient—prices move so quickly when information arises—that no one can buy or sell fast enough to benefit" (*ibid*, p.174). Nonetheless, a variety of patterns have emerged in the past thirty years or so, which apparently seem to contradict the EMH; patterns include low returns on Mondays compared to other days in the week and high returns in January compared to other months in the year. A discussion of these patterns is beyond the scope of this research but a review of several debatable patterns can be found in Malkiel (2007).

Even though methods such as neural networks and genetic algorithms have been used widely for prediction purposes in recent years (see Yoo et al 2005 for a review of some recent developments in stock market prediction models), Hellstrom and Holmstrom (1998) state that the existence of these newer methods does not necessarily contradict the EMH—they say that it just depends on whether or not all interested parties adopt these new methods for prediction purposes. In other words, if

everyone adopts a particular prediction method, then we must assume that no-one has any advantage over anyone else. Even though he supports the EMH, Malkiel (2007) also supports this view – "as more and more people use it, the value of any technique depreciates. No buy or sell signal can be worthwhile if everyone tries to act on it simultaneously" (*ibid*, p.107). To-date, there has been very little research undertaken in the area of automatic event phrase recognition in online financial reports; to that end, our research study describes (1) the development of a prototype financial event phrase (FEP) recogniser that could potentially be used by investors and analysts to automatically analyse online disclosures and (2) the classification of disclosures by likely share price response (up or down) using those FEPs and other features as input.

As our research assumes that publicly available information can have an impact on stock returns, the remainder of this chapter will provide a more detailed overview of the language of financial reports and news (Section 2.3) and the impact of information and events on returns (Section 2.4).

## **2.3 The Language of Financial Reports and News**

Numerous researchers have analysed the language in annual and interim reports and news articles, for various reasons, including narrative analysis (Beattie et al 2004), quantitative and qualitative analysis of content (Back et al 2001; Kloptchenko et al 2004), writing style and tone analysis (Kloptchenko et al 2004; Feldman et al 2008; Loughran et al 2008), thematic analysis (Kohut and Segars 1992), effectiveness analysis (Segars and Kohut 2001), performance and readability analysis (Subramaniam et al 1993; Loughran and McDonald 2011a; Li 2008), positive and negative word analysis (Hildebrandt and Snyder 1981; Thomas 1997), market sentiment analysis (Gillam et al 2002; Ahmad et al 2003; Ahmad et al 2005; Devitt and Ahmad 2007; Daly et al 2009), and summary generation (de Oliveira et al 2002). This section provides an overview of various studies, but the data sources, document features examined, goals, and methods—which were quite varied—are first summarised in Table 2.1. Changes in research goal are highlighted in bold type, to aid scanning. A discussion of the main findings from each of the studies will follow the table.

<i>Author</i>	<i>Page Number</i>	<i>Data Sources</i>	<i>Document Features Examined</i>	<i>Goal</i>	<i>Methods</i>
Beattie et al (2004)	17	Various sections from annual reports.	Text units (phrases).	<b>Narrative analysis to measure the quality of disclosures.</b>	*Developed a four-dimensional framework for the coding/ content analysis of narratives (time orientation, financial/non-financial orientation, quantitative/non-quantitative orientation, and topic and sub-topic categories).
Back et al (2001)	18	Various types of financial documents from the Green Gold Financial Reports' database.	Quantitative and qualitative data (words, sentences, and paragraphs).	<b>Quantitative and qualitative content analysis.</b>	*Used self-organising maps (SOMs) to compare quantitative and qualitative data. *Three SOMs were generated to compare the data: a word map, sentence map, and paragraph map.
Kloptchenko et al (2004)	19	Quarterly reports (online).	Quantitative and qualitative data.	Quantitative and qualitative content analysis. <b>Writing style and tone analysis.</b>	*Used a SOM to analyse quantitative data and prototype matching to analyse qualitative data.
Feldman et al (2008)	19	The Management Discussion and Analysis (MD&A) section of disclosures.	Words.	Writing style and tone analysis.	*Examined changes in tone by counting the frequencies of positive and negative words.
Loughran et al (2008)	20	Form 10-K disclosures.	Terms (phrases).	Writing style and tone analysis.	*Examined ethics-related terms to identify types of firms that tended to use such terms e.g. for 'sin stocks'.
Kohut and Segars (1992)	20	Presidents' letters in annual reports.	Technical characteristics (word count, number of sentences, number of syllables per word, etc.).	<b>Performance and thematic analysis.</b>	*Examined themes and technical characteristics to examine corporate communication strategies. *Independent researchers coded letters on a sentence-by-sentence basis, by theme. *Used stepwise discriminant analysis to determine most discriminatory theme, and statistical methods to identify differences in communication strategies of low- vs. high-performing firms.
Segars and Kohut (2001)	23	CEO's letters to shareholders in annual reports.	Writing style.	<b>Effectiveness analysis.</b>	*Devised measures for evaluating the effectiveness of the letters (credibility, efficacy, commitment, and responsibility).

Subranamiam et al (1993)	23	CEO's letters to shareholders in annual reports.	Writing style.	<b>Performance and readability analysis.</b>	*Examined the readability of the reports.
Li (2008)	23	Annual reports.	Writing style.	Performance and readability analysis.	*Examined the readability of the reports.
Loughran and McDonald (2011a)	23	Form 10-K disclosures.	Technical characteristics (word count, readability).	Performance and readability analysis.	*Computed the readability of 10-Ks using both the Fog index and simple word count. *Used regression analyses to link readability with analyst dispersion and standardised unexpected earnings (SUE).
Hildebrandt and Snyder (1981)	24	Letters to stockholders in annual reports.	Words.	<b>Positive and negative word analysis.</b>	*American and German participants identified positive and negative words. *Participants used these lists of positive and negative words to manually classify letters as positive, negative, or neutral. *Also examined the number of positive and negative words in financially-good and bad years.
Thomas (1997)	25	Management letters in annual reports.	Technical characteristics (passive constructions, verbs, agents, themes, etc.).	Positive and negative word analysis to identify management's motivations and priorities.	*Examined transitivity and thematic structures in the letters by examining the number of passive verb constructions, process verb types, human vs. non-human participants, and themes.
Gillam et al (2002)	27	Reuters news articles.	Words.	<b>Market sentiment analysis.</b>	*Used SystemQuirk to analyse frequency of occurrences of good and bad news terms. *Plotted the good and bad news items as a time series of market sentiment. *Also developed the SATISFI system to extract terminology and to calculate the frequencies of good and bad sentiment words.
Ahmad et al (2003)	28	Reuters news articles (vs. British National Corpus (BNC)).	Phrases and words.	Market sentiment analysis.	*Compared the relative distribution of positive and negative sentiment phrases in the BNC with the Reuters data set. *Used a concordance to examine patterns that preceded and followed sentiment words.

Ahmad et al (2005)	28	Reuters news articles.	Words.	Market sentiment analysis.	*Elaborated on the Gillam et al 2002 sentiment study, to extract sentiment from text using the frequencies of positive and negative words. *Proposed how these methods could be adapted to other areas e.g. to measure the fear of crime.
Devitt and Ahmad (2007)	28	Online news articles about two airline takeovers.	Words.	Market sentiment analysis.	*Explored a metric of sentiment intensity and polarity in text. *Examined the connectivity and position of individual lexical items within a representation of a whole text (lexical cohesion).
Daly et al (2009)	29	Online news stories.	Words (and time series and indices).	Market sentiment analysis.	*Examined correlations between sentiment time series, consumer confidence series, and stock market indices. *Calculated return and volatility for each correlation.
de Oliveira et al (2002)	29	Reuters news articles.	Words.	<b>Summary generation.</b>	*Developed a system called SummariserPort, which uses lexical cohesion to automatically summarise texts. *Research students and traders evaluated the quality of the summaries.

Table 2 .1: D ata s ources, document features examined, g oals, a nd m ethods used i n v a rious s tudies t hat e xamined t he l anguage i n a n nual a nd i nterim reports and news articles.

Beattie et al (2004) outlined five approaches to the analysis of narratives in annual reports and how these approaches might be used to measure the quality of voluntary disclosures. They distinguished between subjective types of analysis (i.e. analysts' ratings) and semi-objective types (i.e. disclosure index studies, thematic content analysis, readability studies, and linguistic analysis). Beattie et al reported that analysts' ratings are no longer used in most countries, as they can be unreliable due to analyst bias. Disclosure index studies involve coding disclosures based on the presence or absence of specific items; the higher the score, the higher the perceived quality of the disclosure. Thematic content analysis involves analysing the whole text for specific themes; themes may be longer or shorter than a sentence and words can

also be used as the measurement unit. Readability studies involve using a formula, such as the Flesch index<sup>8</sup>, to determine the difficulty of the text. Finally, linguistic analysis involves applying numerous indexicals, or criteria for evaluating narratives, to text units or short extracts. For their own contribution, Beattie et al devised a four-dimensional content analysis coding framework, which comprised of time orientation (three categories), financial/non-financial orientation (two categories), quantitative/non-quantitative orientation (six categories), and topic and sub-topic (79 categories). Sentences were split into text units, such that each text unit represented a single piece of information. Their framework presents a method for determining not only the topic but also the three other useful dimensions mentioned above. They also devised a measure of disclosure quality but admitted that their framework is difficult and time consuming to implement.

Back et al (2001) used the Green Global Financial Reports' database and self-organising maps (SOMs) to compare text and numerical data in various financial documents, because both data types are of interest to different stakeholders including auditors, investors, and management. The database consisted of income statements, balance sheets, cash flow statements, financial ratios, and general company information for 160 international pulp and paper companies. They identified 47 key ratios but asked 10 financial analysts from a large Finnish bank to pick the most important variables for the task; this resulted in 9 variables being chosen: operating profit, profit after financial items, return on total assets, return on equity, total liabilities, solidity, current ratio, funds from operations, and investments. They decided not to use all 47 variables as they were concerned they would have insufficient observations and that this would affect the performance of the neural network. The financial data covered the five-year period from 1985 to 1989. This data was then used to create a SOM based on quantitative data.

They scanned in 270 annual reports for 76 of the 160 companies; they did not scan the remaining company reports as they were not available in English and they needed

---

<sup>8</sup> The Flesch Index measures the number of syllables per word and the average number of words per sentence. Each sentence is then assigned a score. The lower the score, the more difficult the text (Flesch 1972).

all qualitative data to be in the same language. They then filtered the annual reports as they only wanted to use US and Canadian reports; this resulted in 234 reports being used to create three SOMs based on qualitative data. The three SOMs – a word map, a sentence map, and a paragraph map – were then used to compare the reports at word, sentence, and paragraph level. Even though their preliminary analysis only concentrated on three companies, they found a "slight tendency" to exaggerate the qualitative text when compared to the actual quantitative state of the company.

Kloptchenko et al (2004) analysed text and financial ratios in quarterly reports downloaded from the Web, with a view to identifying indications to likely future financial performance. They used a SOM to analyse the quantitative data and a prototype matching tool to analyse the text. Using seven quarterly reports for three telecommunication companies during 2000-2001, they found that qualitative and quantitative data seem to represent different things; text tends to give hints about future performance, whereas financial ratios tend to refer to past performance. They also found that the writing style and tone in a company's financial report tends to change before a major company event; the tone tends to represent the future performance more than the past or current performance. However, some limitations of their study include a very small data set (only three companies were used), a limited vocabulary set (they were all telecommunications companies), and a significant usage of proprietary names, all of which may have skewed the results to some extent.

Feldman et al (2008) examined changes in tone in the Management Discussion and Analysis (MD&A) section of disclosures, to see if it adds any incremental information to that already provided by preliminary earnings surprises, accruals, and operating cash flows. By counting the frequencies of positive and negative words, they found that tone changes in MD&As yield excess average returns and that the returns tend to drift for longer periods that extend beyond the subsequent quarter's preliminary earnings announcements. In addition, they found the change in tone was incrementally more informative when firms were small and analyst following was weak.

Loughran et al (2008) examined ethics-related terms in Form 10-K reports between 1994 and 2006, to see if they could identify types of firms that had a tendency to use these terms. They found that, in the pre-regulatory period up to 2002 when the use of ethics-related terms in reports was voluntary, such terms only appeared in 8% of the reports. They also found that terms specifically related to a 'code of conduct' appeared less than 1% of the time. In the post-regulatory period, code-related terms appeared in almost 60% of 10-Ks, as firms were legally obliged to discuss their code. They then focused their attention on reports disclosed during the pre-regulation period to see if any firms that used ethics-related terms had been identified as 'sin stocks'<sup>9</sup>, were involved in class action lawsuits, or had received poor corporate governance scores. They found that these 'problematic' firms were more likely to use ethics-related terms than other firms, probably to appeal to investors who were concerned about deception, but the proportional differences between these 'problematic' firms and other firms were even more pronounced in the pre-regulatory period. They proposed that further study could investigate the link between long-term stock performance and the use of ethics-related terms in 10-Ks, not just for sin stocks but also for poor-governance firms. Whilst they did not report specific results regarding returns, they did report that these sin stocks in their dataset performed "relatively well", unlike the poor-governance firms they studied (p.18).

Kohut and Segars (1992) examined themes within presidents' letters in annual reports with a view to examining corporate communication strategies. The president's letter is most often used to convey organisational strategies to shareholders but it also frequently contains hints as to the implicit beliefs of the organisation. Whilst the final version is usually edited by a public relations or legal officer, the original message usually comes from the president or chief executive officer (CEO). The authors referred to 'direct' as opposed to 'indirect' or circuitous styles of writing—they said that the indirect style "seeks to reduce as much as possible the risk and uncertainty frequently implicit with negative messages" (*ibid*, p.10). They presented three research questions: (1) can technical characteristics of messages (e.g. word count, number of sentences, and syllables per word) differentiate between high-performing

---

<sup>9</sup> Sin stocks typically refer to public companies involved in industries that sell alcohol, tobacco, or gaming products.

and low-performing firms, (2) can the themes presented differentiate high-performing and low-performing firms, and (3) can the time frame of themes differentiate both types of firms?

Their sample comprised the top 25 and the bottom 25 firms of the Fortune 500, based on Return on Equity (ROE). The researchers independently coded 50 letters on a sentence-by-sentence basis; each sentence unit was classified according to its dominant theme. They also categorised each unit on the basis of past or future reference to themes. The reliability of the coding between coders ranged from 92% to 100% and they identified six recurring themes: environmental factors, growth, operating philosophy, product/market mix, unfavourable financial reference, and favourable financial reference (see Table 2.2 for a list of some typical words used to signify the different themes).

<i>Typical words used to signify themes</i>	<i>Words</i>
Environmental factors	<i>recession inflation</i>
Growth	<i>rapidly expanding markets expanding market share</i>
Operating philosophy	<i>building a strong organization strong position</i>
Product/market mix	[general or specific references to the organisation's products and services]
Unfavourable financial reference	<i>short term losses reduction in asset size</i>
Favourable financial reference	<i>gains increased profits</i>

Table 2.2: Typical words used to signify different themes (Kohut and Segars 1992).

They used stepwise discriminant analysis to determine which themes significantly contributed to the discriminatory model and they used discriminant analysis, t-tests, and other statistics to identify differences in communication strategies of high- and low-performing firms.

When they analysed word-count differences between both types of firms, they found that this yielded the only significant difference. The language used by high-

performing firms tended to be much more verbose than low-performing firms. For example, the maximum and minimum word counts in high ROE firms were 4,974 and 366 words whereas the maximum and minimum word counts for low ROE firms were 212 and 187), which suggests that good news is elaborated on and bad news is kept as concise as possible.

When they performed content analyses, one interesting finding was that 85.6% of sentences in high ROE messages referred to past themes as opposed to 75.3% of sentences in low ROE messages. When they examined the specific themes used in past references, they found that operating philosophies were addressed most often for both types of firms (35.8% for high ROE firms and 39.8% for low ROE firms). Operating philosophies were also the most frequently referred-to theme in future references, albeit at a much lower level (5.9% and 17.6% respectively). Another interesting finding was the fact that high ROE companies referenced the past more than low ROE firms and the difference, regardless of theme, was significant between both types of firms. Whilst they found no significant differences between future references, regardless of theme, low ROE firms did tend to reference the future slightly more than high ROE firms, as highlighted above. They suggested that the large number of past references for both types of firms was possibly because executives "feel more confident discussing a certain past rather than an uncertain future" and they want to maintain credibility (Kohut and Segars 1992, p.17).

When they used stepwise discriminant analysis to identify the most discriminatory themes, they found that future references to environmental factors and operating philosophies and past references to product/market mix, unfavourable financial events, and favourable financial events contributed significantly to the model. The model was able to classify 64% of high ROE firms and 92% of low ROE firms correctly or 39/50 firms overall (78% accuracy). They identified some limitations of their study, namely that the 50 firms were selected without considering previous trends, the firms were selected because they belonged to specific industries, and perhaps other measures of success could be used instead of ROE (e.g. sales volume).

Segars and Kohut (2001) devised measures for evaluating the effectiveness of CEO's letters to shareholders, a component of annual reports. They conceptualised an effective letter as one that was successful in achieving credibility, efficacy, commitment, and responsibility. Like Segars and Kohut, Subramiam et al (1993) also examined the CEO's letter to shareholders section of the annual report, but this time with a view to testing the relationship between performance and the readability of the report. They randomly-selected 60 annual reports and found that the annual reports of 'good' performers were easier to read than those of 'poor' performers, because the former tended to use strong, clear, and concise writing unlike the latter who tended to use more jargon and modifiers. Similarly, Li (2008) found that the annual reports of less profitable firms were more difficult to read and firms with reports that were easier to read tended to have more persistent positive earnings.

In a related study, Loughran and McDonald (2011a) found that using the word count of 10-Ks was a better indicator of text informativeness, than using a traditional readability formula such as the Fog index (Gunning 1952). They critiqued the use of Fog index for business documents for several reasons: firstly, it was initially developed to determine the grade level for school textbooks, which is a different readability goal. Secondly, average sentence length (a calculation used by the index) is difficult to calculate accurately in 10-Ks as lengthy sentences are often broken up by bullet points. They also disagreed with using multi-syllable words—also known as 'complex words' by Fog index users—in the readability calculation, as these frequently appear in 10-Ks and yet they are comprehensible to most readers (examples cited included *corporation*, *directors*, and *telecommunications*). They found that 90% of the 10-Ks they examined achieved "unreadable" Fog Index values greater than 20 (values greater than 18 are considered unreadable by anyone without a post-college graduate education) and yet most analysts would probably not consider them unreadable. They decided to use word count, rather than any other measure, for two reasons. Firstly, word count is often used in addition to the Fog Index and is therefore a widely-used measure. Secondly, when they informally asked a sample of partners at accounting firms how they would legally obscure information, the general consensus was that they would bury "the awkward revelation in an overwhelming amount of uninformative text and data" (p.2).

Using regression analyses, Loughran and McDonald used both the 10-K word count and the Fog index score to try to link readability with analyst dispersion (a proxy for uncertainty or divergent opinions regarding a firm's projected earnings) and unexpected earnings surprises (the extent to which the market is surprised). They found that more concise (and therefore more readable) 10-Ks have lower analyst dispersion and standardised unexpected earnings (SUE). They used complete documents in their study. For some related research on 10-Ks, which also involved classification, see Loughran and McDonald (2011b) in Section 3.3.

Hildebrandt and Snyder (1981) applied the 'Pollyanna Hypothesis'<sup>10</sup> to the writing of annual reports and proposed three related hypotheses:

- "Positive words occur more frequently in annual letters to stockholders regardless of a financially good or bad year.
- Negative words occur less frequently in a good year than a bad year.
- German respondents will parallel American respondents when viewing positive and negative words in isolation" (*ibid*, pp.5-6).

They viewed twelve annual letters to stockholders in 1975 (a bad year) and twelve letters in 1977 (a good year). The companies were selected from the Dow Jones Industrials. They used a list of 356 positive and negative antonym pairs and translated these into German. The entire list was given to 100 Graduate Business students, who identified the preferred (positive) and non-preferred (negative) words; this was done to make sure there was a agreement. They then read the 24 annual letters, recorded occurrences of each of the 356 words, and then placed each occurrence into one of three classifications (positive statement, negative statement, or neutral statement), depending on its contextual usage. With regard to the first and second hypotheses, they found that there were significantly more occurrences of positive words than neutral or negative words, regardless of year i.e. 68.9% of words were positive in context in 1975 (a bad year) and 79.5% of words were positive in context in 1977 (a good year). We would expect this to be the case as companies will

---

<sup>10</sup> In the context of business communication, the Pollyanna Hypothesis states that people tend to use positive words more frequently and directly than negative words (Hildebrandt and Snyder 1981).

obviously prefer to use positive language than negative language, whenever possible. These results were found to be statistically significant using t-tests. They proved the third hypothesis by showing the words to the American and German reviewers in isolation (i.e. not in context) and found that over half of the preferred or positive words selected had 90% agreement in both languages. It is important to note that when these words were used in context, the meaning often changed quite significantly. For example, whilst *increased* is a preferred word and *decreased* a non-preferred word, in the following statement, *increased* has a negative connotation: "The OPEC nations, because of their consolidated position, were able to *increase* crude oil prices about \$1 per barrel in October" (*ibid*, p.9).

Thomas (1997) examined the differences between good news and bad news communicated in management letters at the start of annual reports, with a view to identifying management's motivations and priorities. To avoid differing company styles, she gathered the annual reports for one company over a five-year period (1984-1988). She chose this particular period as the company in question changed from being a profitable to an unprofitable one during that period. Her main research goal was to confirm or question the 'Pollyanna hypothesis' in the context of annual reports, as discussed in Hildebrandt and Snyder (1981).

Thomas examined transitivity structures and thematic structures. Transitivity "describes a clause according to the kind of verb used, the participants, and the circumstances" (*ibid*, p. 53). With regards transitivity structures, she first looked at the number of passive verb constructions in each clause in the management letters. She found there was an increase in the number of passive constructions from 10% in 1984 (a profitable year) to 20% in 1988 (a loss-making year). Thomas suggested that this makes sense, as messengers tend to distance themselves from negative messages whenever possible. Distancing is a feature of deception and can be examined by investigating linguistic styles (Newman et al 2003). Using Halliday's (1985) six process verb types (material, mental, relational, behavioral, verbal, and existential), Thomas found that relational process verbs were used frequently; they appeared in 30-48% of clauses over the five-year period. Relational process verbs are verbs of attribution or identification that convey meanings such as 'being', 'identifying', and

'attributing'. She also found that verbs of 'being' in the first and last paragraphs doubled in the five-year period from 33% to 66%, giving a much more objective or factual impression that should not be questioned. Typically, the first paragraph prepares the reader for the message that will follow and the last paragraph reminds the reader of the key points, so these are both important paragraphs. There was also an increase in the number of non-human participants acting as agents (examples cited include *opportunities*, *fiscal 1988*, and *machine tool market*). Focusing again on the first and last paragraphs, Thomas found that there was a shift away from the writer of the message; instead, nonhuman circumstances were causing these factual situations, so they could not be questioned. Nonhuman participants only appeared in 25% of clauses in 1984 (a profitable year) but they appeared in 87% and 73% in 1987 and 1988 respectively (these were both loss-making years).

"Thematic structure in systemic theory is the part of the clause that serves as a point of departure for the message - what the message is about" (Thomas 1997, p.56). When Thomas examined thematic structures, she found that there were two types of themes: (1) the personal pronoun *we* and (2) a variety of inanimate nominal groups, such as the nonhuman agents mentioned previously. Even though the usage did not progressively increase or decrease each year, she found that overall usage of the pronoun *we* decreased from 75% in 1984 to 27% in 1988 and the usage of inanimate nominal groups increased from 25% in 1984 to 73% in 1988. These thematic findings are largely in line with the transitivity findings outlined earlier, suggesting that management distanced itself from the poor performance in the latter years. Expanding on thematic structures to include an analysis of the discourse practices of the genre, Thomas examined context and cohesion and found that management were happy to take credit for the early successes as they were due to effective leadership; on the other hand, they blamed nonhuman participants for the subsequent problems, and followed these comments up with suggestions that the situation would improve – words such as *nevertheless* and *however* were often used. Thomas also briefly examined condensations; one example of a condensation in Business English is *transition*, as in *transition year*, which could mean things are improving if linked with recent profits, or disimproving if linked with losses or declining profits. Another example is *gradually improving market* – does this mean that the situation was poor

in previous times and now it is improving or that the situation is not improving as well as had been hoped? Does *profitability* mean that they (only) have sufficient funds to pay a dividend or that they are making a healthy profit? In one of the 1987 reports, she identified the following sentence in the first paragraph: "*Cross & Tracker had a difficult year in fiscal 1987 as markets proved weaker than expected and pricing pressures and other factors combined to severely erode margins.*" (*ibid*, p.63). This sentence, which contains several contractions (*markets proved weaker, pricing pressures, other factors, and erode margins*), demonstrates that once again, management clearly abdicated responsibility for the dire financial situation.

Gillam et al (2002) extracted market sentiment from Reuters' news articles to see if there was any correlation between share prices and news announcements. Their Reuters corpus consisted of 30 news items per day for the month of November 2001 as well as the closing values of the FTSE 100 for each day. They used their own SystemQuirk system to analyse the frequency of occurrences of over 70 terms that conveyed 'good' news and over 70 terms that conveyed 'bad' news; these terms were identified in previous related research (Ahmad et al 2002). Using the term frequencies and the closing values, Gillam et al were able to plot roughly the good and bad news items as a time series of market sentiment. They also developed a prototype system called SATISFI, which was able to handle XML-based news, to extract terminology, and to calculate the frequencies of good and bad sentiment words.

As this research was a preliminary investigation into possible correlations between news and FTSE closing values, Gillam et al did not report many specific results. However, some interesting findings include the fact that bad news seems to anti-correlate with the FTSE i.e. they found that two days after bad news was released, the FTSE rose. They evaluated their prototype system using 172 news stories for one week in February 2002. Six native English speakers were asked to determine whether they felt a particular news item would have a good or bad impact on the stock market. Their SATISFI system also performed the same task. They found that the native speakers were better able to correlate good and bad news items with the FTSE. SATISFI was able to correlate bad news but to a lesser extent and it was not able to

correlate good news. Some reasons cited for the problems with SATISFI included the fact that some 'entities' in news items may have different effects on other entities i.e. an event that may be good for one reason, may be bad for another reason or it may have opposite knock-on effects, depending on the company's position at the time. Other problems include negation and double negation, which completely alter the meaning of good and bad news items, and words that are more indicative of good and bad news than others e.g. accounting fraud is most certainly a bad news item but restructuring could be considered a good or bad item, depending on the context.

In a follow-on study, Ahmad et al (2003) compared the relative distribution of positive and negative sentiment phrases in the British National Corpus (BNC)<sup>11</sup> with a Reuters data set of over ten million tokens published between 2000 and 2002. They found that words such as *up* and *down* were more predominant in the BNC than in the Reuters data set. For example, *up* appeared in 91.8% of the BNC but in only 57.2% of the Reuters data. On the other hand, they found that words such as *growth* and *rose* increased significantly in the Reuters collection. For example, the word *growth* appeared in only 5.66% of the BNC but in 25.9% of the Reuters data. Using a concordance, they further examined the patterns that preceded and followed key sentiment words in the Reuters data. The phrase *rose X percent* appeared in over 58% of the patterns that contained the word *rose*. Further studies showed that for the 'rise' and 'fall' sentiment words, the frequencies of related words or synonyms were much lower. For example, variations of *rise* (*rose*, *rising*, or *risen*) occurred in 1,004 patterns; *jump* only occurred in 133 patterns, *climb* only occurred in 75 patterns, and *lift* only occurred in 33 patterns.

Other related studies that examined the language of financial texts include Ahmad et al (2005) who outlined how the qualitative (textual) and quantitative (time series) methods used in their FINGRID project could be adapted to other areas e.g. to analyse online crime statistics to measure the fear of crime or to measure ethnic/racial tension in a community. Devitt and Ahmad (2007) explored a metric of sentiment intensity and polarity in text with a view to comparing it to human judgements of

---

<sup>11</sup> <http://www.natcorp.ox.ac.uk/>

polarity in financial news. Daly et al (2009) examined correlations between sentiment time series, consumer confidence series and stock market indices, calculating return and volatility for each correlation. Their preliminary findings suggested that the returns for sentiment time series, in particular, deviate from that of normal distribution of a random variable. de Oliveira et al (2002) developed a system (SummariserPort) which uses lexical cohesion to automatically summarise financial texts.

Sections 2.1 and 2.2 outlined the on-going conflicting debates surrounding the EMH and the two types of investment analysis. This section (2.3) discussed a selection of studies which examined the language of financial reports and news. As outlined in Table 2.1 at the start of this section, research goals varied considerably, from narrative analysis (e.g. Beattie et al 2004) to performance and readability analysis (e.g. Li 2008). Studies examined words only (e.g. Feldman et al 2008), terms or phrases (e.g. Loughran et al 2008), technical characteristics of the document (e.g. Loughran and McDonald 2011a) and overall writing style (e.g. Segars and Kohut 2001). Methods also varied considerably—for example, self-organising maps (Bach et al 2001), coding frameworks (Beattie et al 2004), readability formulae (Subramaniam et al 1993), and word frequency analyses were used (Gillam et al 2002). As we have now established that information in financial texts can be analysed and used for various purposes, the next section will report on various studies which outlined the relationships, if any, between information, financial events, and market reactions. Each of these sections are relevant to our research project as we assume that the language used to describe some financial events, in certain publicly available documents, can have an impact on the value of financial instruments.

## **2.4 Information, Events and Market Reactions**

Various studies have examined the impact that public news stories and events can have on stock returns and there is significant disagreement about the impact they can have. As outlined in Sections 2.1 and 2.2, proponents of the EMH argue that the market reacts immediately to news and events and therefore they have little or no impact on returns; fundamental and technical analysts, who disagree with the EMH,

argue that news and events *do* impact on the market. As we assume that there is a correlation between some events and stock prices (and therefore believe that the market is not fully efficient), it is necessary to explore further some of the literature in this area.

Ball and Brown (1968) examined the usefulness of accounting income numbers in terms of their information content and timeliness. In terms of information content, they found that one half or more of all the information that is made available about a particular firm during the year is captured in the income number for that year. In terms of timeliness, however, the annual report fell short, mainly because other forms of media—such as interim Form 8-K reports—tend to release that information beforehand. They suggested a need for further study into the impact that the magnitude, not just the sign, of any unexpected income changes could have on stock prices. Our study looks at Form 8-K disclosures as companies are legally obliged to file them within a short timeframe, whenever material events arise (see Chapter 4 for a discussion). We argue that 8-Ks are timelier sources of information than Form 10-K annual reports.

Fama et al (1969) examined the relationship between returns and stock split announcements. They found that there were often unusually high returns in the months preceding a stock split, even though the information about the impending split was not yet known by the investing public. They proposed that this was probably because the company had experienced particularly high earnings and dividends in the months preceding the split and therefore the market re-evaluated its expectations of the company. They argued that the market's judgements regarding the implications of the split were reflected in the price at least by the end of the month when the split was announced but possibly even immediately after the announcement was made; this conclusion supported the EMH.

In a later paper, Fama (1991) loosened his definition of the EMH stating "on average, stock prices seem to adjust within a day to event announcements" (*ibid*, p.1601). In our study, we use a three-day window (days  $t-1$  to  $t+1$ , where  $t$  is the filing date of the Form 8-K), as we expect the market to react to material events within a day. Fama

summarised some of the main findings from event studies in corporate finance, as follows:

- Positive or negative dividend changes tend to be associated with stock price changes of the same sign.
- Stock prices react unfavourably to new issues of common stock.
- In the case of merger and tender offers, stockholders of target firms tend to benefit from large gains.
- In the case of proxy fights, management buyouts and other corporate control events, target stockholders tend to benefit greatly.

In his 1998 paper, Fama discussed long-term return anomalies which had started to emerge in recent literature on event studies, and which seemed to contradict the EMH. He argued that, in an efficient market, many market over-reactions would be balanced out by under-reactions. Secondly, he argued that many long-term anomalies tended to vary depending on the models or statistical methodology used to measure them, and therefore successful results should just be attributed to chance. In our research, we suspect that anomalies that have nothing to do with the arrival of new information, might partially explain some of our results—the identification of these anomalies is beyond the scope of our research however.

Antweiler and Frank (2006) commented on Fama's (1998) complaint that only successful event studies or studies which examined single event types were being published: “*Splashy results get more attention, and this creates an incentive to find them*” (Fama 1998, p.287). Rather than focus on single event types, we examine multiple event types. Antweiler and Frank also examined multiple event types but they considered the returns day-by-day for over a month, whereas we used a relatively short three-day window. Also, they examined Wall Street Journal news stories, whereas we used Form 8-Ks (see Section 3.3 for a more detailed discussion of the Antweiler and Frank classification studies). Antweiler and Frank also commented on the abundance of literature on event studies that has emerged in recent years; as a result, the remainder of this section will examine a selection of the literature in this area.

Cutler et al (1989) examined whether market movements during the 1987 crash were attributable to factors other than the arrival of information. They found that they were only able to attribute about one third of the variance in stock returns to the arrival of macroeconomic information. They analysed stock market reactions to 49 non-economic events from 1941 to 1987; events included the attack on Pearl Harbour, the death of Roosevelt, the Cuban missile crisis, and the Chernobyl meltdown. When the events were aligned with S&P percentage changes, some of the events, including the attack on Pearl Harbour, seemed to be clearly associated with the changes. However, for the set of 49 events, the average absolute market move was only 1.46% compared with 0.56% over the entire period, which demonstrates the "surprisingly small effect of non-economic news" (p.5). They then examined the largest changes in daily returns between 1946 and 1987 and tried to correlate those changes with New York Times explanations. For many of the significant changes, they were unable to identify fundamental causes.

Wysocki (1999) examined the effects of message-posting volume on stock message boards and found that changes in daily posting volume were associated with news announcements relating to *earnings* and *changes in stock trading volume and returns*. He also found that message-posting volume could be used to predict changes in stock trading volume and returns the following day and that certain stocks, particularly those that previously yielded "extreme past performance", yielded the highest volume of postings (*ibid*, p.4). Tumarkin and Whitelaw (2001) also looked at message boards and found that, for Internet-related stocks, returns and trading volume following abnormal message activity were not statistically significant. However, they also found that on days with abnormally high message activity, strong changes in opinion correlated with abnormal returns and abnormally high trading volume, and the abnormally high trading volume persisted for a second day. Overall, however, their use of event studies and auto regression lead them to the conclusion that there is no causal link between message board activity and stock returns or trading volume.

Stice (1991) examined price and volume reactions to *earnings* announcements to see if the reactions were caused by the release of SEC filings or by the subsequent announcements of same in the Wall Street Journal (WSJ). Using a dataset of

predominantly small firms, he did not find any significant reaction to the initial SEC filings, perhaps because of the limited dissemination possibilities<sup>12</sup>. However, he did find a reaction to the subsequent WSJ announcements and the increase/decrease in price change was consistent with the increase/decrease in unexpected earnings.

Like Stice, Abarbanell and Bernard (1992) also examined stock price responses to *earnings* announcements but with a view to clarifying whether or not analysts' forecasts under-react or over-react to such announcements<sup>13</sup>. With regard to under-reactions, they found that stock prices tended to respond to earnings news with even more delay than analysts and suggested that this might be because of transaction costs or because traders under-react to the analysts' under-reactions. With regards to over-reactions, they could not clearly link analysts' over-reactions with stock price over-reactions following earnings announcements. Finally, they proposed that the disagreement amongst researchers about under- and over-reactions arose because studies documenting over-reactions used prior stock price performance to partition firms whereas other studies partitioned using prior earnings performance.

Skinner (1994) looked at voluntary disclosure of positive and negative *earnings* surprises. As one might expect, he found that managers voluntarily disclose positive earnings surprises to present the company in a positive light. More notably, he found that managers frequently pre-empt the possible impact of disclosing large negative earnings surprises, by voluntarily disclosing the bad news before they are obliged to do so. He also found that bad news disclosures have a greater impact on price than good news disclosures and that they are more likely to contain qualitative statements about quarterly, as opposed to annual, earnings-per-share.

---

<sup>12</sup> At the time of this study, members of the public could only access SEC filings by photocopying them at an SEC public reference room or by hiring another firm to source the data for them. Both methods tended to prove costly, financially and time-wise (Stice 1991). For many people, the Wall Street Journal was more accessible. Since 1995, however, SEC filings have been made freely available online on EDGAR, so dissemination time has been greatly reduced.

<sup>13</sup> By under-reaction, we mean that the average return following a good news announcement is higher than the average return following a bad news announcement. An over-reaction arises when the average return following a *series* of good news announcements is lower than the average return following a *series* of bad news announcements. See Abarbanell and Bernard (1992) for an overview of some of the literature on analyst under- and over-reaction.

Other researchers who examined under- and over-reactions to news include Barberis et al (1998), Daniel et al (1998) and Tetlock (2008). In their model of investor sentiment, Barberis et al highlighted two phenomena from behavioural psychology—representativeness and conservatism. Representativeness can be defined as the tendency of interested parties to view events as typical and to ignore the laws of probability (e.g. the fact that a company that has been growing for quite a while might not always keep growing at that rate). Conservatism, which tends to be closely tied in with under-reaction, is defined as the inability of interested parties to update opinions even when presented with new data or evidence (Edwards et al 1968). They developed a model of investor sentiment which proposed that *earning* announcements and other similar *corporate events* represent information that has low strength but significant statistical weight. Their model predicted that such events would result in a stock price under-reaction. On the other hand, they found that consistent patterns of news, such as series of good or bad earnings announcements, were of high strength but low statistical weight, and would probably result in stock price over-reaction. They recommended further studies into the apparent strength of news events.

Behaviouralists believe that market prices are highly imprecise and that investors are irrational<sup>14</sup>; they also believe there are four factors that create irrational market behaviour—overconfidence, biased judgements, herd mentality, and loss aversion (Malkiel 2007). Daniel et al examined two of these factors—overconfidence and biased self-attribution. They defined an overconfident investor as "one who overestimates the precision of his private information signal, but not of information signals publicly received by all" (*ibid*, p.1841). Biased self-attribution means that the investor's confidence tends to grow when public information is aligned with his/her private signals but it does not fall commensurately when they are not aligned. Daniel et al found that stock prices over-react to private information signals and under-react to public signals. In Appendix A of their paper, they present some relevant literature about under-reaction to public news events. Events cited include *stock splits, tender offer repurchases, open market repurchases, analyst recommendations, dividend initiations and omissions, seasoned issues of common stock, earnings surprises,*

---

<sup>14</sup> See Barberis and Thaler (2003) for a review of behavioural finance.

*public announcement of previous insider traders, venture capital share distributions, and earnings forecasts.* Advocates of the EMH (see, for example, Fama 1970 and Malkiel 2007 in Section 2.2) would argue that these anomalies are merely chance deviations; others, including Daniel et al would argue that the returns are too strong and regular to be chance deviations so there must be some behavioural model which can explain the anomalies. In our study, we assume that the market reacts to some event types within a short timeframe but we do not try to prove or disprove the EMH; rather, we assume that the market is not fully efficient.

Tetlock (2008) examined market overreaction to stale information i.e. information that has previously been released elsewhere. He found that returns partially reverse after stale news is released as individual investors overreact to the stale information. Likewise, he found that news that has not previously been released tends to yield much smaller return reversals or else yields a continuation of the positive returns.

A number of studies looked at specific event types (Fair 2000; Fair 2003; Liu 2000; Ng and Fu 2003). Fair identified 69 macroeconomic events that led to large changes in S&P 500 futures prices. For the purposes of his study, one-to-five minute changes greater than or equal to 0.75% in absolute value were deemed large changes. He searched the Dow Jones News Service, Associated Press Newswire, New York Times, and Wall Street Journal sources for news reports by time of day. There were some price changes for which no event could be found; in these instances, Fair argued that brokers would most likely say there was no event or perhaps there was simply renewed confidence in the market. 22 of the 69 events related to *money supply, interest rate announcements, or testimony by monetary authorities*; 14 events related to *employment reports*, and the remainder concerned various other *macroeconomic* issues. In a subsequent study, Fair (2003) identified 152 additional events that could lead to a similar change in various futures—S&P 500 futures, bond futures, Deutsche Mark futures, Yen futures, and British futures. In this study, however, Fair's goal was to determine if a systemic relationship actually existed between the changes and stock prices, bond prices, and exchange rates. Also, the new collection of 221 events comprised not only of macroeconomic news but also microeconomic news; 12 categories of news were identified ranging from *US fiscal policy* events to

*international conflicts*. The four news sources that were used in Fair (2000) were used in this study but for a longer time period.

Liu (2000) used a sample of 611 *innovation news announcements* made by 103 US biotechnology firms to determine investors' behaviour after the announcement and to determine if investors were able to interpret the information in an unbiased manner. The announcements were obtained from the Lexis/Nexis database and spanned the period from 1983 to 1992. The Lexis/Nexis database includes sources such as Business Wire, PR Newswire, and Reuters. Liu used five categories of innovation news announcements – *FDA approval*, *patents granted*, *scientific breakthroughs*, *strategic alliances or research joint ventures*, and other *technology-related news*. Liu intentionally excluded bad news announcements for a number of reasons, namely (1) bad innovation news was much less frequent so it would have been difficult to gather a reasonable sample of such news, (2) firms do not always disclose bad news so it would have been difficult to be subjective about selecting suitable announcements, and (3) good news announcements about innovations tend to have a great deal of uncertainty about them and Liu wanted to see how the market reacts to uncertain announcements. Overall, Liu found that investors tend to be over-optimistic about innovation news and therefore tend to misprice innovative events, particularly with firms that are not very research and development (R&D) intensive, firms that are large, and firms that have high book-to-market ratios. Eventually, investors tend to revert to examining firm-specific fundamentals when attempting to value a given event.

Ng and Fu (2003) examined event episodes, which they defined as a set of event types within a specific window of time. Examples of event types include *Cheung Kong stock goes up* and *US increases interest rate*. In particular, they examined frequent stock episodes i.e. many instances of the same episode within a specific window. They used a tree structure, or event tree, to represent the event types and then pruned the tree based on event type frequencies, to reduce the likelihood of including non-frequent episodes. Each window was counted at most once, even if the event type appeared in several days during the window. They used a recursive mining procedure and then evaluated the performance using synthetic and real data.

The real data was extracted from a repository of local newspapers and contained 121 event types spanning 757 days. The stock data was retrieved from the Datastream International Electronic Database for 12 local companies. They did not report many findings but they did suggest a relationship between the NASDAQ and PCCW telecom stock (i.e. both went down).

Other studies investigated positive and negative news or sentiment (Hong et al 2000; Conrad et al 2002; Li 2006; Lerman and Livnat 2009). Hong et al (2000) examined the existence of momentum in stock returns and found that negative firm-specific news tended to diffuse more slowly across the investing public, than positive news. They defined momentum as the situation whereby "past winners continue to perform well, and past losers continue to perform poorly" (*ibid*, p. 265). They considered analyst coverage as a proxy for the rate of information flow on the assumption that, all else being equal, stocks with lower analyst coverage should take longer to diffuse information across the investing public. To counteract the likelihood that larger companies tend to have more analyst coverage, they performed a regression of coverage on firm size to produce the residual analyst coverage.

Conrad et al (2002) examined behavioural and regime-shifting models<sup>15</sup> to determine if stock prices respond more strongly to bad news than good news when relative market valuations are rising. They also wanted to determine if both types of models needed to be extended to take firm-specific reactions as well as market conditions and valuations into account. They used a sample of 24,097 announcements of firms' annual earnings between 1988 and 1998 and measured the level of the market by comparing the market price-earnings (*P/E*) ratio during each firm-announcement month with the monthly average market *P/E* during the preceding 12 months. A high *P/E* month, for example, meant that the market's forecast of *P/E* was higher that month than it was in previous months. Their findings broadly supported the hypothesis that stock prices respond more strongly to bad news than good news when prices are already high.

---

<sup>15</sup> With behavioural models, investors employ past firm-specific information when forming expectations and they ignore aggregate market conditions; with regime-shifting models, investors consider market-wide information and they do not consider firm-specific news.

Li (2006) counted the frequency of risk and uncertainty words in Form 10-K reports, to determine the impact of risk sentiment on future earnings and returns. Using regressions, he found that firms with a large increase in risk sentiment led to more negative earnings changes and significantly negative returns the following year, compared to firms with a smaller increase in risk sentiment.

Lerman and Livnat (2009) examined market reactions to Form 8-K disclosures (the subject of our research) to see if these timely disclosures diminished the impact of the annual Form 10-K and quarterly Form 10-Q reports and also to determine the market reaction, if any, to 8-Ks<sup>16</sup>. They found that the impact of 10-Ks and 10-Qs was not diminished despite the arrival of 8-Ks, possibly because investors and analysts use 10-Ks and 10-Qs to interpret the effects of material events previously disclosed in 8-Ks. In fact, they found that the information content of 10-Ks and 10-Qs, as measured by abnormal trading volume and abnormal return volatility, increased once these new regulations were introduced.

With regard to market reaction, they found that reactions varied by event—some event items caused strong positive mean abnormal returns, others caused negative returns. For event items that did not cause any abnormal returns, they suggested that there may be an absence of information or inconsistent reactions (e.g. for some firms, a particular event item may be good news, for others it may be bad). The reaction may also vary depending on the company's position at that time. They also examined abnormal trading volume and abnormal stock return volatility and found that there were significant reactions to all event items when they used these measures, which also suggests that events that may be good for one company may be bad for another. Finally, they found that the market tends to under-react to 8-Ks and that the significant return tends to drift for some event items.

---

<sup>16</sup> See Lerman and Livnat for a review of some other studies which examined market reactions to Form 8-Ks or to other reports that contain events which must now be disclosed in Form 8-Ks. The majority of these studies, which were prior to 2004, examined market reactions to earnings announcements. As the filing regulations have since changed, it is not clear if the Form 8-Ks disclosing the earnings news were filed the same day as the reported reaction.

## 2.5 Summary

In this chapter, we outlined the on-going debates amongst efficient market advocates and investment analysts regarding how the market reacts to information and news. There is still no clear consensus. We also looked at various studies that examined the language of financial reports and news and learned that there is a slight tendency to exaggerate qualitative text when compared to the actual quantitative state of the company (Back et al 2001), tone changes can signify impending events and/ or yield excess average returns (Kloptchenko et al 2004; Feldman et al 2008), and ethics- and code-related terms are used more by certain types of companies (Loughran et al 2008). There were conflicting reports on the style of language that tends to be used by good-/poor-performing companies, in terms of readability, occurrences of positive/negative words, and occurrences of passive constructions (Hildebrandt and Snyder 1981; Kohut and Segars 1992; Subramaniam et al 1993; Thomas 1997). We also looked at the relationship between information, financial events, and market reactions. Event studies have found reactions to some news events but reactions varied greatly and depended on the data and techniques used. These studies demonstrate that financial language can be complex and varied and it is therefore difficult to identify clear-cut patterns that can be correlated with market reactions.

Whilst it is evident that there is no clear consensus about how *consistently* the market reacts to financial news and events<sup>17</sup>, there is still plenty of scope for content analysis of financial events in various sources (e.g. online news stories and SEC reports) by examining various features (e.g. single words, phrases, or financial variables), using various techniques (e.g. statistical or automatic analysis). The goals of these analyses can vary from prediction of returns to identification of market trends and sentiment.

---

<sup>17</sup> Malkiel (2007), who supports the EMH, stated: "No one can consistently predict the direction of the stock market or the relative attractiveness of individual stocks, and thus no one can consistently obtain better overall returns than the market. And while there are undoubtedly profitable trading opportunities that occasionally appear, these are quickly wiped out once they become known. No one person or institution has yet to produce a long-term consistent record of finding money-making, risk-adjusted individual stock trading opportunities, particularly if they pay taxes and incur transaction costs" (*ibid*, p.246).

Kroha and Baeza-Yates (2004) argued that because experts interpret news in different ways and markets react differently and in different timeframes to different news, it is not possible to create a training collection that is correlated with guaranteed market responses. They argued that information retrieval was not strong enough to identify the difference between positive and negative news and that more sophisticated language models and analyses were needed if one wishes to use news as an investment tool. Whilst we agree that it is not possible to create a *definitive* training collection, we do believe there is some *potential* in using linguistic methods to recognise events and keywords in disclosures and in using these features to predict the likely share price response. Whilst one would not be able to *consistently* predict the likely share price, our methods could be used to assist in the arduous task of analysing and classifying responses manually.

In the past twenty years or so, various studies have used automatic techniques to analyse the content of financial news (Swales and Yoon 1992; Kryzanowski et al 1993; Wüthrich et al 1998; Cho et al 1999; Qi 1999; Lavrenko et al 2000; Thomas and Sycara 2000; Das and Chen 2001; Gidófalvi 2001; Peramunetilleke and Wong 2002; Thomas 2003; Antweiler and Frank 2004; Koh and Low 2004; Koppel and Shtrimberg 2004; Kroha and Baeza-Yates 2004; Lam 2004; Seo et al 2004; van Bunningen 2004; Fung et al 2005; Antweiler and Frank 2006; Kroha et al 2006; Liu et al 2006; Mittermayer and Knolmayer 2006a; Schumaker and Chen 2006; Tetlock 2007; Engelberg 2008; Henry 2008; Tetlock et al 2008; Loughran and McDonald 2011b). As each of these studies used automatic content analysis techniques and are therefore more closely related to our research, they are discussed in detail in Chapter 3.

# Chapter 3: Automatic Analysis and Classification of Financial Documents

## 3.1 Outline

In the previous chapter, we outlined the two sides to the efficient markets/investment analysis debate and found that there is no clear consensus. We also looked at various studies which examined the language of financial reports and news and examined the relationship between information, financial events, and market reactions. We learned that financial language can vary greatly depending on the events encountered and the 'state' of the company. Event studies demonstrated reactions to some event types but not to others, and these reactions varied depending on the techniques and data used.

In this chapter, we will briefly introduce the topics of automatic text classification and categorisation (Section 3.2) and then we will look specifically at studies that used automatic techniques to analyse the textual content of *financial* documents, as we also used automatic techniques for this purpose (Sections 3.3 to 3.5). Whilst *some* of these studies also involved automatic classification as an end-goal, we will use the term 'automatic analysis' to describe all the studies, regardless of their end-goal. We do not discuss literature on financial time series prediction, as that is beyond the scope of our research. However, for a good review of various methods that can be used for this purpose, see, for example, Tay et al (2003).

Section 3.3 will describe studies that used single words, and possibly other features, for the automatic content analysis of financial documents (Das and Chen 2001; Swales and Yoon 1992; Schumaker and Chen 2006; Mittermayer and Knolmayer 2006a; Fung et al 2005; Antweiler and Frank 2004; Antweiler and Frank 2006; Gidófalvi 2001; Liu et al 2006; Lavrenko et al 2000; Koppel and Shtrimerberg 2004; Kroha and Baeza-Yates 2004; Kroha et al 2006; Tetlock 2007; Tetlock et al 2008; Engelberg 2008; Henry 2008; Loughran and McDonald 2011b). Whenever authors used other features as well as single words, all these features are discussed together in this section for completeness.

Section 3.4 describes studies that used keyword records and phrases, but not single words (Peramunetilleke and Wong 2002; Wüthrich et al 1998; Thomas and Sycara 2000; Cho et al 1999; Thomas 2003; Seo et al 2004; van Bunningen 2004). In these studies, keyword records contain more than one keyword. As mentioned previously, studies that used keyword records and phrases as well as single words (Das and Chen 2001; Swales and Yoon 1992; Schumaker and Chen 2006; Mittermayer and Knolmayer 2006a) are only discussed in Section 3.2.

Section 3.5 describes studies that used financial ratios and variables (Koh and Low 2004; Qi 1999; Kryzanowski et al 1993; Lam 2004). Only one study used financial ratios or variables with another feature (Lavrenko et al 2000). Because this study also involved the analysis of single words, it is only discussed in Section 3.2.

Finally, Section 3.6 will summarise the key findings from the chapter.

## **3.2 Automatic Text Classification and Categorisation**

Lewis (1992b) states that there are two main groups of content-based text processing tasks: text classification and text understanding. Text classification involves “the assigning of documents or parts of documents to one or more of a number of groups” (*ibid*, p.2). Text understanding, on the other hand, is a more complicated process that may involve “extracting formatted data, answering questions and summarization or abstracting” (*ibid*, p.2).

Text classification, the focus of our research, can be further broken down into a number of different activities—document retrieval, text retrieval, text categorisation, text routing, term categorisation, document clustering, term clustering, and latent indexing (Lewis 1992b). Our research focuses on text categorisation, which is the “classification of documents with respect to a set of one or more pre-existing categories” (*ibid*, p.3). Whilst the most common application of text categorisation is indexing documents for text retrieval with a view to creating document representatives, it can also be used for text understanding to filter out uninteresting sections of data, or to categorise documents directly for a human user. Our research

focuses on this latter type of text categorisation—specifically, we aim to predict the likely share price response (up or down) based on the textual content of financial documents.

The remainder of this section will briefly introduce topics related to text classification, including document and text retrieval, as well as machine learning. In Sections 3.3 to 3.5, we will discuss the topics of automatic financial analysis and classification in detail, as these are most relevant to our own research.

In his text ‘Information Retrieval’, van Rijsbergen (1979) discussed early work on automatic text analysis that focused on statistical rather than linguistic approaches. Luhn (1958), who focused on the automatic creation of literature abstracts, examined word frequency and distribution, stating that “the frequency of word occurrence in an article furnishes a useful measure of word significance” (*ibid*, p.160). Obviously, word frequency alone should not be the sole consideration—hence the introduction of stop-lists and other techniques to eliminate any “useless” frequently-occurring words. Luhn also went on to say that words can be considered more significant if they are frequently found near each other within the same sentence and that the relative position of these significant words *within* a sentence, is a useful determiner of significant sentences. It is important to note here that such measures of significance do not consider relevant linguistic issues such as grammar and syntax (*ibid*).

According to Sparck Jones and Willett (1997), “the task of an IR system is to retrieve documents or texts with information content that is relevant to a user’s information need.” (*ibid*, p.1) As stated earlier, document retrieval is essentially a classification task (Sparck Jones and Willett 1997; Lewis 1992b) as it is undertaken for a specific purpose (van Rijsbergen 1979). Document retrieval comprises two main activities: indexing and searching. In the early research, automated searching was the main focus, but later automated indexing became a priority also (Sparck Jones and Willett 1997). Indexing is undertaken to classify documents so they match future requests and searching is undertaken to classify a file into “matching and nonmatching parts” (*ibid*, p.2).

van Rijsbergen highlighted three basic steps that usually need to be taken when developing a text processing system that can generate a document representative from an input text. An input text could be a title, abstract, or the complete text. A document representative is needed so it can be used by an automatic retrieval system and it can be generated using statistical, linguistic, or knowledge-based methods [Lewis 1998]. The three basic text processing steps are as follows:

1. Remove high frequency words using a stoplist.
2. Strip suffixes using a complete list of suffixes and remove the longest one. One major disadvantage to this method is that it ignores context, so context rules should ideally be devised and suffixes only removed, if the context is right. van Rijsbergen cites the examples of FACTUAL and EQUAL, where one would not want to remove the suffix UAL from EQUAL.
3. Detect equivalent stems. Once suffixes have been removed, the creation of a list of equivalent stem-endings can further help the process. Two stems should only match if they have equivalent stem-endings e.g. BANDPT for ABSORB- and ABSORPT-. These stems should be combined into one, or conflated. However, words with equivalent stem endings are not necessary equivalent. For example, NEUTRON and NEUTRALISE should not be indexed as one concept. Whilst errors are inevitable, the goals should be to minimise the errors as much as possible.

The document representative that emerges from the text processing process can contain features such as single keywords, n-grams, phrases, terms, or named entities (Lewis 1998; Kosala and Blockeel 2000). Commonly used feature selection methods include the bag-of-words approach, information gain, Term Frequency by Inverse Document Frequency (TF\*IDF), and the vector space model (Fagan 1987; Kosala and Blockeel 2000).

Salton (1970) highlighted Luhn's contributions stating that newer indexing and text analysis systems typically comprise the following types of operations:

1. Words, word stems, noun phrases, and other "content units" are chosen from the user's query text (*ibid*, p.335).

2. Weights are assigned to each content unit (or expression), typically based on the frequency, position, or type of expression. The commonly-used bag-of-words approach to indexing (Caropreso et al 2001) represents a document as a vector of weights based on the frequency of each word in the document. Words (features) are assumed to be independent of each other i.e. word order is not important (Kosala and Blockeel 2000; Manning et al 2009). Other variants of the bag-of-words approach use word stems (as opposed words) or simply the presence or absence of words; this latter variant is also known as binary feature weighting (Lewis 1998) or the set-of-words approach (Kosala and Blockeel 2000; Caropreso et al 2001).
3. Expressions can be developed further using a stored dictionary or by considering statistical measures of co-occurrence between terms in a collection, or syntactical relations between terms.
4. Each document to be indexed is identified by a set of weighted terms, not all of which necessarily appear in the document.

Early indexing experiments compared automatically-derived index terms with a list of manually-derived topics (Salton 1970). Human experts typically developed the performance evaluation criteria in these early systems. Later studies included Cleverdon's Cranfield experiments, which involved manual indexing but automated searching and Salton's SMART project, which automated indexing as well as searching (Salton 1970; Sparck Jones and Willett 1997). Salton's project, in particular, extended the early text analysis methods to documentation in various fields. Also, these systems defined many of the effectiveness measures that are still used today, including recall and precision (Salton 1970; Sparck Jones and Willett 1997). Recall measures the proportion of relevant content retrieved. Precision measures the proportion of retrieved material that is *actually* relevant (Salton 1970; Frakes and Baeza-Yates 1992). Whilst these figures are plotted against each other and the goal is to achieve a high score for both—1 is the ideal—Salton reminds us that broadening the search formulation can lead to high recall but narrowing the search formulation can lead to high precision (see Figure 3.1). Oftentimes, it is necessary to compromise, by aiming for a high recall rate but only an acceptable precision rate. Salton recommended using interactive search methods to improve

retrieval performance. These methods use multiple searches based on user feedback information supplied during the search process.

More recently, IR researchers have turned their attention to using non-statistical methods for indexing, in particular those methods that focus on semantically richer data such as syntactic phrases or a combination of phrases and words (Caropreso et al 2001). However, one disadvantage to using phrases on their own is that they must occur frequently enough to have an impact; also, several apparently different phrases can mean the same thing (Caropreso et al 2001).

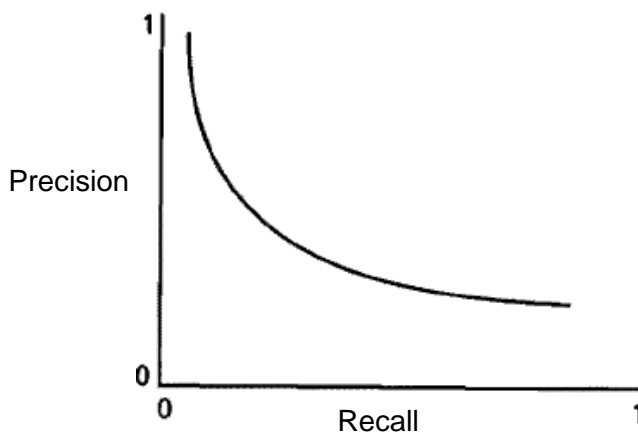


Figure 3.1: Recall-precision graph (Frakes and Baeza-Yates 1992 p.11)

We will now turn our focus to automated text classification, the subject of our research. Sparck Jones and Willett state that classifications can be static or dynamic, the latter “involving learning in response to training examples or feedback” (*ibid*, p.2). Machine learning includes both supervised and unsupervised learning methods (Lewis 1992b). Unsupervised learning methods include word or document clustering as well as term weighting methods, which aim to find structure in large bodies of text that can improve document representations. Supervised learning, on the other hand, is the process of “mechanically building a classifier for a set of known classes” (Lewis 1992b p.11). Our research involves supervised learning.

We discussed earlier that there are a variety of textual representations that can be used to analyse text. Likewise, there are a number of different types of machine learning algorithms and methods that have been used extensively for various purposes,

including data mining (Quinlan 1993; Mitchell 1997; Tay et al 2003). Machine learning methods include:

- Decision trees/ rule induction classification models.
- Support vector machines.
- Neural networks.
- Genetic algorithms.
- Bayesian classifiers.
- Instance-based learning methods.
- Statistical machine learning techniques.

Our study represents documents using automatically-extracted financial event phrases and keywords and we experiment with two classification methods—decision trees and support vector machines. We also compare the performance of these experiments with two commonly-used baseline approaches to document representation—bag-of-words and n-gram representation [Kossala and Blockeel 2000]. In the bag-of-words experiments, we use Naïve Bayes for classification but in the n-gram experiments, we use decision trees and a support vector machine. A discussion of each of these experiments can be found in Chapter 6.

As the literature in the area of automated text classification is vast, we now concentrate our literature review on the automatic analysis and classification of *financial* documents, a specific focus of our research. A number of different data sources, features, goals, and methods have been used for this purpose. Table 3.1 outlines the data sources used (online news stories or messages; on- or off-line statements or reports) and the features examined (single words; keyword records and phrases; financial ratios and variables). Some authors used combinations of features, whilst others just used one feature type. In the studies described in this chapter, the financial ratios and variables are extracted from document content, not from time series. Goals also varied (prediction of prices, prediction of direction or trends; identification of market or message sentiment; other goals). 'Other goals' include predicting going concern companies and predicting the volatility of companies or markets. A number of different machine learning methods have also been used; the most frequently used methods include decision trees and rule induction

algorithms, neural networks, support vector machines, and Bayesian methods. However, statistical methods, language modelling, multiple discriminant analysis, k-nearest neighbour, and genetic algorithms were also used. As these latter methods were not used very often, they have been categorised together as 'other methods' in Table 3.1. In Chapter 6, we will discuss the features of the two machine learning algorithms we used—C4.5 and SVM-Light (Sections 6.2.1 and 6.3.1 respectively).

Author(s)**	Page Number	Data Sources		Document Features Examined			Goals			Methods*				
		Online news stories or messages	On- or off-line statements or reports	Single words	Keyword records and phrases	Financial ratios and variables	Prediction of prices, directions, or trends	Identification of market or message sentiment	Other goals	Decision trees and rule induction methods	Neural network methods	Support vector machines	Bayesian methods	Other methods
Das and Chen (2001)	50	*		*	*			*					*	*
Swales and Yoon (1992)	53		*	*	*		*				*			*
Schumaker and Chen (2006)	54	*		*	*		*				*			*
Mittermayer and Knolmayer (2006a)	56	*		*	*		*				*			*
Fung et al (2005)	60	*		*			*				*			*
Antweiler and Frank (2004)	61	*		*				*			*	*		
Antweiler and Frank (2006)	62	*		*			*					*		
Gidófalvi (2001)	63	*		*			*					*	*	
Liu et al (2006)	65	*		*			*							*
Lavrenko et al (2000)	67	*		*		*	*					*	*	
Koppel and Shtrimerberg (2004)	69	*		*				*		*		*	*	
Kroha and Baeza-Yates (2004)	71	*		*			*					*	*	
Kroha et al (2006)	73	*		*								*	*	
Tetlock (2007)	74	*		*			*							*
Tetlock et al (2008)	75	*		*			*							*
Engelberg (2008)	76	*		*			*							*
Henry (2008)	79	*		*		*	*							*
Loughran and McDonald (2011b)	80		*	*			*							*
Permutilleke and Wong (2002)	85	*			*		*		*	*				*
Wüthrich et al (1998)	87	*			*		*		*	*				*
Thomas and Sycara (2000)	89	*			*		*							*
Cho et al (1999)	90	*			*		*							*
Thomas (2003)	92	*			*		*		*	*				
Seo et al (2004)	94	*			*		*					*	*	
van Bunnigen (2004)	96	*			*		*				*			*
Koh and Low (2004)	98		*			*		*	*	*				*
Qi (1999)***	99	?	?			*	*			*				*
Kryzanowski et al (1993)	100		*			*	*			*				
Lam (2004)	101		*			*	*			*				
<b>Slattery (2012)</b>			*	*	*		*			*		*	*	

Table 3.1 Data sources, document features examined, goals, and methods used for the automatic analysis and classification of financial documents.

\*The categories of methods have been adapted from Berthold and Hand (2003) and Tay et al (2003).

\*\*The papers are listed in the order in which they are discussed (i.e. by features examined).

\*\*\*Qi (1999) did not provide information on the content sources used.

### 3.3 Single Words

Das and Chen (2001) experimented with five different algorithms to measure emotive (not factual) sentiments in Yahoo message postings and to classify messages as bullish (optimistic), bearish (pessimistic), or neutral<sup>18</sup>. Neutral messages included spam messages and ambiguous messages that were deemed neither bullish nor bearish.

A number of pre-processing techniques were employed. Firstly, the messages were downloaded from the Web using a Web scraper program, which downloaded thousands of messages from Yahoo in a few minutes. The HTML tags were then parsed out of the text and all abbreviations were expanded. Finally, they performed some grammatical processing in order to tag negation effects in sentences. The purpose was to ensure the classifier could process sentences that contained reverse meanings (e.g. the statement *Company X is reporting a profit this year* means the opposite to *Company X is not reporting a profit this year*).

Their approach used three sources of data. Firstly, the Computer User Version of the Oxford Advanced Learner's Dictionary (CUVOALD) was used for basic lexical items (e.g. adjectives, adverbs, and parts-of-speech tagging information). They also used a lexicon of financial words. The words in the lexicon were hand-selected but also statistically processed, to determine the strongest discriminators (i.e. the words that were most likely to help the classifier). Each word in the lexicon was presented with its base value (a figure which corresponded to the classification category it would most typically belong to e.g. 0 for null messages, 1 for sell messages, and 2 for buy messages). Each lexical entry also contained all forms of the word, expanded abbreviations, and negation effects. The third source of data was the grammar, which typically comprised a training set of 300 to 500 manually classified messages; as mentioned previously, each message was classified as bullish, bearish, or neutral. It is important to note that there was a disagreement rate of 27.54% when they asked a second reader to classify manually the same 374 training messages and 64 test messages. This demonstrates how ambiguous online messages can be, and why

---

<sup>18</sup> A bullish investor is eager to buy or sell and a bearish investor is too shy to buy or sell.

manual or automatic classification of such messages can be extremely difficult.

The entire collection comprised over 25,000 messages relating to eight stocks, spanning approximately 100 stock days. The stock prices at the time of each message posting were also downloaded. They employed five algorithms: naïve Bayes (NB), a vector distance classifier (VDC), a discriminant-based classifier (DBC), an adjective-adverb phrase classifier (AAPC), and a Bayesian classifier (BC). They also experimented with three voting schemes. The first voting scheme required three of the five classifiers to agree on the message type (bullish, bearish, or neutral). The second scheme required four of the classifiers to agree, and the third scheme required all five classifiers to have consensus on the message type. Even though the accuracy of the classification improved significantly when they moved from simple majority to consensus, the number of messages classified decreased. Because two humans hand-tagged the training messages and they disagreed on several messages, the authors employed voting schemes to ensure that only messages that could be agreed on by the majority (or all) of the schemes were used during classification, thereby guaranteeing high levels of precision but low levels of recall. Ambiguous messages were deemed not useful and were therefore not classified by the algorithms.

As mentioned previously, the two human subjects who hand-tagged the training messages disagreed on 27.54% of the classifications, which gave an "agreement coefficient" of 72.46%. Das and Chen used the agreement coefficient as one of the performance benchmarks; the other benchmarks were perfect performance (100% classification accuracy) and random guessing (33% in this three-way classification). The five classifiers were each evaluated for their individual classification performances (attempted classifications only, ignoring any ambiguous messages) and then the voting schemes were employed, to see if they improved the performances.

With the individual algorithms, the best performance was with the BC, which achieved 53.13% accuracy on the 64 test messages. The AAPC achieved a promising result also; 51.56% of the 64 test messages were correctly classified, implying that an adjective or adverb coupled with the two words immediately following or preceding it in the message provides a useful indication of sentiment. The poorest performance

was with the VDC, which achieved 42.19% accuracy on the 64 test messages.

When Das and Chen employed the simple majority voting scheme (three of the five algorithms agreed), 61.54% accuracy was achieved when 80% of the test messages were attempted; this 80% attempt rate compares favourably with the 72.46% agreement coefficient. The classification accuracy increased to 85% when four algorithms agreed but the number of attempted messages decreased to 31.25%. When there was consensus amongst all five algorithms, the classification accuracy rose to 100% but only 7.81% of the messages were attempted.

Das and Chen also created a sentiment index to measure the sentiments in the messages. For each message, they created a time series of sentiment, by incrementing the sentiment index by 1 whenever a bullish message was posted (i.e. bullish according to the algorithms) and decrementing it by 1 whenever a bearish message was posted. The index began at 0 and the stock prices were collected whenever a message was posted. In most cases, the visual plots of the index<sup>19</sup> clearly showed a correlation between stock prices and sentiment. In some instances, the sentiment responded almost immediately to the stock price change; in other instances, the sentiment lead the stock price, and in other instances, the sentiment lagged the stock price. A subsequent correlation analysis of lead and lag times, found that for the most part, sentiment lagged stock price by approximately 50 minutes (i.e. it took about 50 minutes for message board sentiment to reflect the stock price) but there were also some instances when the sentiment lead the stock price. This prompted them to undertake a more in-depth phase-lag analysis to identify whether a major pattern first emerged in the stock graph or the sentiment graph. They found that in almost all cases, the sentiment lagged the stock price so the message sentiment was not a useful predictor. In a related study, Das et al (2005) also found that stock returns drive message sentiment. In terms of limitations, the researchers did not compare their methods to baseline approaches and they did not adopt a trading strategy.

---

<sup>19</sup> See Figures 2-7 in the Appendices of Das and Chen (2001).

In an earlier study, Swales and Yoon (1992) examined the ability of artificial neural networks (ANNs) to predict stocks that are likely to perform poorly or well, based on the contents of the President's Letter to Stockholders, a component of the Annual Report. The performances of different ANNs were compared with the performance of multiple discriminant analysis (MDA).

Swales and Yoon created two data sets using the Fortune 500 listings and Business Week listings. The 58 companies with the highest returns were selected from the former source and the 40 companies with the highest market valuations were selected from the latter source. They then split these two datasets into two sub-groups: the highest and lowest returns or highest and lowest valuations respectively. The Business Week data was used as the training set and the Fortune 500 data was used as the test set.

For each company in each dataset, both the supervised neural network and a nine-variable MDA used the frequency of references to common themes to predict the future stock price performance. The percentage of letters that alluded to each particular theme was also taken into account. Themes were identified using content analysis of similar words and phrases and they included references to growth, strategic plans, and anticipated gains and losses. Unfortunately, the authors did not state whether the content analysis was performed manually or automatically.

Swales and Yoon tested three different ANN architectures: a two-layer, three-layer, and four-layer neural network. They experimented with different architectures to see if the hidden units in the three- and four-layer ANNs supported any non-linear function between the input (the theme of the letters) and output units (are they good or poor performing firms?). They did not state how many neurons there were per layer. When a three-layer or four-layer architecture was used, the ANN performed better than the MDA with the training set. For example, the three- and four-layer ANNs both achieved 86% accuracy when attempting to classify the high valuation companies in the training set, whereas the MDA only achieved 72% accuracy. When attempting to classify the low valuation companies, the four-layer ANN significantly outperformed the MDA, achieving 96% accuracy, as opposed to 76% for the MDA.

For the same low valuation companies, the three-layer ANN performed only marginally better than the MDA, achieving 79% accuracy as opposed to 76%.

The test data produced slightly different results. The three- and four-layer ANNs outperformed the MDA for high return companies (they both achieved 90% accuracy, as opposed to 70% for the MDA) but the results for low return companies were much lower for the ANNs and much closer to the MDA results. The three- and four-layer ANNs achieved only 50% and 65% accuracy respectively, both much closer to the MDA accuracy (60%). Looking at just the mean results for the testing data, the four-layer ANN achieved the highest mean accuracy (77%), followed by the three-layer ANN (70%), and then the MDA (65%). The two-layer ANN achieved the lowest mean accuracy with the testing data (52%).

The authors also had another interesting finding; during the training phase, the MDA, two-layer ANN, and four-layer ANN were more accurate at predicting low valuation companies, than high valuation companies. However, during the testing phase, the MDA and the three ANNs were more accurate when predicting high return companies, than low return companies. Unfortunately they did not state the learning rates or propose any rationale for these unusual results. Also, they did not report on the statistical significance of any of their findings and the only evaluation metric they used was prediction accuracy.

Schumaker and Chen (2006) experimented with three textual representations of Yahoo! Finance online news articles, to determine which representation can best predict future stock prices. Unlike most other authors, their goal was to predict discrete stock prices, not the direction of the stock price, although they did evaluate the directional accuracy also. The three representations were bag-of-words, noun phrases, and named entities. They gathered 9,211 articles and over ten million stock quotes for a five-week period during 2005. They decided to ignore any articles that occurred within 20 minutes of another article, to eliminate articles that might lead to conflicting results. They then filtered out any terms that occurred less than three times in any document, which resulted in the following breakdown for each textual representation (see Table 3.2).

<i>Textual Representation</i>	<i>Number of Terms</i>	<i>Number of Articles</i>
<i>Bag-of-words</i>	5,285 terms	2,853 articles
<i>Noun phrases</i>	5,293 terms	2,861 articles
<i>Named entities</i>	2,858 terms	2,760 articles

Table 3.2: Breakdown for each textual representation (Schumaker and Chen 2006).

Their prototype system (AZFinText) used the three representations for each news article and a derivative of a support vector machine (SVM) to predict the stock price 20 minutes after the release of the article. They used three evaluation metrics for the SVM and compared the results of each metric with the equivalent results from linear regression (LR). The first metric was closeness, which used the mean squared error to compare the predicted stock price with the actual stock price. The second metric was directional accuracy, which compared the predicted direction of the stock price change with the actual stock price direction. The third metric was a simulated trading engine, which evaluated each article and bought (shorted) the stock if the predicted +20 minute stock price was greater than or equal to 1% movement from the price when the article was released. After 20 minutes elapsed, any bought/shorted stocks were sold. Their SVM model outperformed the LR for all three evaluation metrics.

Looking at the simulated trading engine results for SVM and LR respectively, they found that the bag-of-words approach would have earned \$5,111 (SVM) versus a loss of \$1,809 (LR). Similarly, noun phrases would have earned \$6,353 (SVM) versus a loss of \$1,809 (LR). Finally, named entities would have gained \$3,893 (SVM) versus a loss of \$1,879 (LR). The authors also examined each textual representation, to determine which one was best at predicting stock prices. Interestingly, some representations performed better than others, depending on the metric used. For example, named entities had the lowest mean squared error (MSE) when measuring closeness (0.03346), bag-of-words had the next lowest (0.04713), and noun phrases had the highest MSE (0.05826) with p-values < 0.01. However, when they measured directional accuracy, named entities had the poorest result (49.2%) and noun phrases had the best result (50.7%).

Schumaker and Chen examined these apparently conflicting results further, by looking at the cash outlay required to make these returns. Even though noun phrases apparently yielded the highest profit (\$6,353), the simulated trading strategy for noun phrases required an investment of \$295,000, yielding a return of 2.15%. They proposed that even though noun phrases were good for directional accuracy, their investment strategy was inadequate for prediction purposes due to the large number of poor stocks chosen. Named entities, on the other hand, which apparently yielded the lowest profit (\$3,893), actually yielded the highest return (3.60%) as an investment of only \$108,000 was needed. In other words, a high directional accuracy does not necessarily imply a highly profitable trading strategy. Schumaker and Chen proposed that named entities generate a “minimally representative essence” of news articles and that these “essences” are better than noun phrases for short-term prediction. They highlighted some limitations of their approach, namely that their dataset was relatively small and they only used companies from the Standard and Poors (S&P) 500; perhaps specific industry groups would have yielded better results. They also proposed that other machine learning techniques could be evaluated. Whilst they did not consider transaction costs, inflation, or spread, particular strengths of this study include their use of various evaluation metrics (baseline approaches, prediction accuracy, closeness, and a simulated trading strategy) and their use of 10-fold cross validation and statistical measures to ensure robustness.

Mittermayer and Knolmayer (2006a) examined the impact of PRNewswire press releases on stock price trends. Like other authors who assumed that the markets react promptly to news and events (see Section 2.4 for a discussion on how markets react), Mittermayer and Knolmayer decided to use intra-day prices rather than end-of-day prices. They developed a hand-made thesaurus of words and phrases that they proposed had the largest impact on prices; the thesaurus consisted of single words (e.g. *up* and *down*), phrases (e.g. *formal investigation* and *sales climb*), and tuples of words and phrases (e.g. *approve* near the words *financial guidance*). The thesaurus was used in conjunction with a bag-of-words model to define the features present in each press release. However, if a press release contained a feature that was not present in the bag-of-words but was present in the thesaurus, then that feature was still included in the training data.

Their NewsCATS system contained three sub-components: a document pre-processing engine, a categorisation engine, and a trading engine. The document pre-processing engine created the feature list for the training data (press releases) using the bag-of-words method and various feature selection functions, including inverse document frequency (IDF), collection term frequency (CTF), and information gain (IG). IDF was the default function but the other settings were also evaluated. The number of features was limited to 15% of the number of documents in the training collection, using the 10 to 15 features recommended by Lewis (1992b) as a guideline. The engine then mapped the training data into vectors, using the values obtained from various functions—within-document frequency (WDF), IDF, WDF\*IDF, and Boolification. The categorisation engine then used these vectors with various algorithms to categorise each press release into a category. The algorithms included a linear support vector machine (SVM), a non-linear SVM, k-nearest neighbour (k-NN), and Rocchio. Finally, the trading engine recommended a trading signal based on the categorisation results.

The authors extracted transaction prices from the NYSE Trade and Quote Database. Prices for the S&P 500 were in 15-second intervals over a nine-month period in 2002. They then calculated a series of two-minute moving averages after the press releases were made available resulting in 49 moving averages, which were compared with the average one minute before and the average one minute after the press release was made available. Whilst they initially obtained 18,000 press releases from PRNewswire, they then performed filtering to remove releases posted at weekends, releases posted outside of normal trading hours, and releases that referred to more than one ticker symbol; this filtering resulted in 9,128 press releases. The press releases were then filtered again to select only those they felt could actually have an impact on share prices. This final filtering resulted in 989 press releases that referred to one or more of seven pre-defined topic classes. The seven topic classes were 'dividends', 'earnings projections or forecasts', 'financing agreements', 'legal issues', 'licensing/market agreements', 'offerings', and 'sales reports'. We believe this final level of filtering was excessively restrictive as only 989 of the initial dataset of 18,000 press releases were used. A system should be capable of dealing with all types of news, not just those that were deemed by the researchers to have an impact.

The 989 press releases were made up of 83 'good', 42 'bad', 504 'neutral', and 360 'unclear' press releases. Unlike the previously discussed authors who used either two or three prediction categories, the authors introduced a fourth category (unclear) to catch press releases that were ambiguous and could potentially create noise in the system. Unclear releases were omitted from training but they were used for testing purposes. For unexplained reasons, nine releases were randomly excluded, resulting in a final collection of 980 releases.

Using 90% of the up, down, and neutral collection for training and 10% of the up, down, and neutral collection for testing purposes, they performed 10-fold cross validation for all the experiments. Of the remaining 360 unclear releases, which were not used for training, 10% of these were added to each of the ten test sets.

Rather than implement a one-sided early exit strategy which would require one to sell if the investment rose by a certain percentage or more, they implemented a two-sided early exit strategy. In their strategy, if the investment rose 0.5%, they capitalised on the gain by selling; on the other hand, if the investment declined by 2% or more, they implemented what they called a "stop loss transaction" (*ibid*, p. 3). In other words, whenever 'good' predictions were made, they bought the stock long and sold later and whenever 'bad' predictions were made, they sold the stock short and bought back later. For all their experiments, they evaluated the results with and without this two-sided early exit strategy.

The average performance of NewsCATS using the IDF default setting was first compared with random guessing. Looking at the harmonic mean of macro average precision and recall, NewsCATS achieved 66% accuracy compared with 33% for random guessing. Looking at the overall accuracy (i.e. the percentage of correct predictions), NewsCATS achieved 82% accuracy, compared with 33% for random guessing. When they did not use their early exit strategy, they achieved a profit per roundtrip of 0.22%, marginally lower than the 0.23% profit achieved by Lavrenko et al. (2000). One roundtrip occurs when an investor purchases a quantity of stock in one transaction and sells that quantity of stock the same day, regardless of how many sale transactions are involved; two roundtrips occur when an investor makes two

separate stock purchases the same day and sells them all the same day. However, when they implemented the early exit strategy, the NewsCATS profit increased to 0.27%; they assume that their careful selection and filtering of data to eliminate noise contributed to this improved result.

Mittermayer and Kolmayer performed a robustness analysis on each of the adjustable parameters by varying the feature selection function, the size of the feature set, the document representation, and the classifier. The best-performing feature selection function was CTF, which achieved a harmonic mean of macro-averaged precision and recall of 69% and an overall accuracy of 83%. Note how these results are marginally higher than the previously reported results achieved by IDF (66% and 82%). CTF also earned the highest profit per roundtrip (0.28%) and the highest overall profit (94%), when the early-exit strategy was employed. When they varied the size of the feature set between 5% and 25% of the number of documents, they did not find any improvement in performance over the default setting of 15%. Whilst the profit per roundtrip with an early-exit strategy was highest when the feature set was only 5% of the document collection (0.34%), the total profit was the lowest overall (88%) as there was a smaller number of roundtrips overall. When they tried various techniques for representing the document vectors, they found that the default method (WDF\*IDF) performed the best; the profit per roundtrip was 0.28% and the total profit was 94%, when an early-exit strategy was used. The IDF and Boolean representations resulted in profits per roundtrip below zero. When they varied the algorithms, they found that a non-linear SVM with a polynomial kernel earned the highest overall profit with and without the exit strategy (91% and 98% respectively). However, the harmonic mean of macro-averaged precision and recall was only marginally higher (68.9%) than linear SVM, the default algorithm (68.7%).

Whilst the authors did report a high profit per roundtrip with an early exit strategy (0.29% when they used non-linear SVM), they did not take transaction costs into account. They stated that this was intentional, as they wanted to compare their study with previous studies. They also did not consider inflation or costs associated with the spread. Some strengths of this study include a range of evaluation metrics (random guessing, prediction accuracy, precision and recall, and adoption of a trading

strategy), their use of 10-fold cross-validation, and automated feature selection methods (including IDF).

Fung et al (2005) focused on aligning the words in 350,000 Reuters news stories with tertiary movements in stock prices using the term frequency-inverted document frequency (TF\*IDF) weighting scheme and the Efficient Markets Hypothesis (see Section 2.2 for a discussion). Whilst they defined a tertiary movement as a movement that "lasts for less than three weeks, which denotes the short-term market behaviour" (p.1), Faerber (2000) defined tertiary movements as daily movements that are "inconsequential due to their erratic nature and volatility" (p.181).

Fung et al warned against the danger of incorrectly aligning a news story with the general trend of a time series; for example, even though the general trend may be rising, the exact observation when the news story is released (if using the EMH formulation) could be a decrease. To counteract the problem of incorrect alignment, they used a segmentation algorithm. Each document in the training collection was then represented by a Vector Space Model and classified using a support vector machine (SVM) algorithm into one of three categories: positive, negative, or neutral, depending on whether the stock price rose significantly, dropped significantly, or did not rise or drop. The breakdown of training and testing data was not provided and it is not clear if cross validation was used.

They described two simulation methods for evaluating the performance of their prototype system: (1) they employed their prototype system, which requires one to buy and sell frequently, depending on the content of news stories and (2) they employed the buy-and-hold test, which requires one to buy and hold everything at the start and not sell anything until the end. They calculated the rate of return for both methods and found that the first far outperformed the second (the accumulated return results were 18.06 and -20.56 respectively). Whilst they assumed zero transaction costs and they did not consider inflation or costs associated with the spread, they found that the rate of return for their prototype was found to be statistically significant at the 0.5% level (versus a randomised test), when they undertook 1000 additional simulations.

They also performed a hit-rate analysis to determine how often the hit rate (sign of the return) could be predicted accurately. They found that their prototype system yielded the highest hit rate and accumulated return within a prediction period of three-to-five days. For a prediction period of three days, the hit rate was 61.6% and the accumulated return was 18.06. For a prediction period of five days, the hit rate was 65.4% and the return was 21.49. Strengths of this study include their use of various evaluation metrics (random guessing, hit rate analysis, and adoption of a trading strategy) and their use of statistical analysis to verify robustness.

In another prediction study, Antweiler and Frank (2004) looked at the content of over 1.5 million Yahoo! Finance and Raging Bull message board postings as well as Wall Street Journal news entries for 45 companies. The messages were typically 20-50 words in length and one message was downloaded per day for each company.

In their first experiment, they used 1,000 messages that were manually classified as 'buy', 'hold', or 'sell' (25.2%, 69.3%, and 5.5% of the collection respectively). 'Hold' messages should not be bought or sold. They restricted the vocabulary set to the top 1,000 words, as ranked by information gain, and then used naïve Bayes to classify those training messages. Even though they also experimented with a support vector machine (SVM) algorithm, they did not report on those results, as they were similar to those of naïve Bayes, i.e. 18.1% ('buy'), 65.9% ('hold'), and 4.1% ('sell') respectively.

The authors also classified the complete collection of over 1.5 million messages but they did not report on the accuracy of those classifications, possibly because they had only manually classified 1,000 messages (although this was not stated). Only the total percentages of 'buy', 'hold', or 'sell' classifications were given; unfortunately, these figures cannot be directly compared with the matrix of manual classifications, as they may include misclassifications. Other findings include an "economically small but statistically robust" relationship (*ibid*, p. 1261) between the number of postings and negative subsequent stock returns the subsequent day, and a strong positive relationship between the number of message postings and volatility.

In a follow-on study, Antweiler and Frank (2006) changed their focus to classifying Wall Street Journal news stories by topic, rather than classifying message board postings as bullish or bearish. Their study differed from many previous event studies in that they examined the impact of multiple event types rather than single event types, thereby avoiding publication bias. They also used various event windows that spanned more than a month around the news event date; windows started three days before the event to three days after and were 5, 10, 20, and 40 days in duration. They read a random sample of news stories and identified 67 basic news topics and 7 possible qualifiers. When they restricted the collection to topics which had at least 50 stories, this resulted in 43 categories. Like their previous study, they restricted the vocabulary set to the top 1,000 words, as ranked by information gain, and then used naïve Bayes to classify over 200,000 news stories by topic.

With regard to the topic classifications, they found that many of the classification results matched conventional expectations. See Table 3.3 for some examples using an event window that commenced two days before the news release.

<i>News Topic</i>	<i>5-Day Response (t-2 to t+3)</i>	<i>Statistical Significance Level</i>
<i>Earnings reported up</i>	5.2	99.9%
<i>Earnings reported down</i>	-14.7	99.9%
<i>Dividend increased</i>	20.0	99.9%
<i>Dividend decreased</i>	-21.2	95.0%
<i>Stock split</i>	33.4	99.9%

Table 3.3: Average standardized cumulative abnormal returns for news events, scaled by 100 for readability purposes (Antweiler and Frank 2006).

They found that the length and starting day of the window can have a significant impact on the analysis of returns. For example, when they considered a 5-day event window, all the windows that started before the news was released (i.e.  $t-3$  to  $t-1$ , where  $t$  is the release date) had statistically significant positive abnormal returns. However, when they considered a 20-day window, the three starting dates had statistically significant negative returns. Therefore, depending on the window used, one could incorrectly make the assumption that the majority of the news was good (or

bad). They also found that statistically significant returns generally became stronger as the event window was extended to include more post-event days i.e. when the information was in the public domain; this scenario appears to reject the EMH<sup>20</sup> (see Section 2.2 for a discussion). They found that the initial strong reaction to news, regardless of whether it was pre- or post- the official news release date, was typically followed by a more lengthy reversal or over-reaction. They also found that the temporary jump in trading volume which accompanied the initial jumps, tended to be followed by a gradual decline in trading volume. Finally, they found that the average news stories tended to have a bigger and more prolonged impact during recessions than during expansion periods. Strengths of this study include their use of feature selection methods (information gain) as well as multiple windows and statistical measures to ensure robustness. However, evaluation of their results was limited only to classification accuracy.

Gidófalvi (2001) used a naïve Bayes classifier to learn numerical indicators from Biz Yahoo! news stories. The experimental dataset comprised 12 stocks with high-frequency intra-day prices over a three-month period between November 1999 and February 2000. Any articles that fell outside normal trading hours or outside the time intervals, for which they had stock prices, were disregarded. Depending on the alignment window used (to be discussed shortly), there were 4,300-4,650 articles in the training set and 1,300-1,650 articles in the test set.

Gidófalvi aligned each article with its stock price using a window of influence with upper and lower time boundaries (e.g. [0,20], where 20 is 20 minutes after the release of the article) and then scored each article using the  $\beta$ -value. The  $\beta$ -value is a measure that compared the movement of the stock with the arithmetic average of the selected stock prices. A movement of zero implied the stock price movement was as expected; a movement greater than or less than zero implied that the movement was respectively better or worse than expected. It is important to note here that some stocks that received a negative movement score actually had positive stock price changes during the window of influence, and vice versa. This is because the  $\beta$ -value

---

<sup>20</sup> Antweiler and Frank did point out, however, that their rejection of the EMH might be attributable to aspects of market microstructure, such as the length of time it takes traders to unwind large portfolio positions.

measure is a relative measure. It is based not only on the stock price change, but also on the index price change and the *expectation* regarding how the stock price might react to this change. When they calculated the  $\beta$ -values for individual stocks using linear regression, they found that some of the regression results had very low  $r^2$  values; in other words, the predicted value did not correctly model the actual change of the stock price. The author cited one possible reason for this, namely that several articles might contain important news but only the first article might actually influence the stock price.

Using these measures, he labelled each article as having an 'up', 'down', or 'expected' movement. Like Antweiler and Frank (2004), Gidófalvi restricted the vocabulary set to the top 1,000 words, as ranked by information gain, and then used naïve Bayes to classify those training messages. Unfortunately, the feature words were not provided in this paper. However, in a subsequent paper, which used a different dataset but built on this study, Gidófalvi and Elkan (2003) provided the first 100 words. Words like *incumbent*, *exploited*, *unsophisticated*, *damn*, and *detergents* were deemed highly predictive for the up class and words like *counted*, *fang*, *indefeasible*, *upgradeable*, and *quadruples* were deemed highly predictive for the down class. Some of these predictive words are quite unusual and as Mittermayer and Knolmayer (2006b) pointed out, it is difficult to accept that the top five words (*sbc*, *msft*, *websphere*, *db*, and *index*) could be useful predictors.

Gidófalvi (2001) experimented with different labelling threshold values to ensure that articles that contain negative words (e.g. *shortfall* or *bankruptcy*) would be correctly labelled as 'down' and not as 'expected'; likewise, he wanted to ensure that articles that contain positive words (e.g. *propel* or *peak*) would be correctly labelled 'up' and not as 'expected'. When he examined the prediction accuracy and its statistical significance, the best classification (relative to random guessing) and the highest statistical significance, were achieved when the negative threshold value was  $-0.002$  and the positive threshold value was  $+0.002$ . These values were then used for all classification experiments.

Gidófalvi also evaluated different alignments, to determine if the information in each article influenced the stock price before the news article was released to the public, or after it was released. The best prediction accuracy, relative to random guessing, was achieved when alignments 20 minutes prior to an article being released and 20 minutes after an article being released, were used. Precision and recall were also greatest for these alignments. For alignments greater than  $\pm 20$  minutes, the classification accuracy decreased. Even though the overall predictive power of the classifier was low, he reported that the alignment result indicated that there is some correlation between news articles and stock prices. Some strengths of this study include their use of multiple evaluation metrics (random guessing, prediction accuracy, and precision and recall), and statistics to measure robustness. In the follow-on study, Gidófalvi and Elkan (2003) implemented a minimal risk trading strategy and found that for the  $[-20,0]$  and  $[0,+20]$  alignments, the naïve Bayes classifier performed significantly better than 1000 random traders for the prediction certainty threshold  $\delta=0.33$ .

Liu et al (2006) used Yahoo! Finance message board postings to identify top expert posters, to assign weights to their predictions, and to generate profitable returns. In their study, "experts" were defined as any posters who contributed information or opinions on the message boards. Because these experts could potentially post inaccurate predictions up to 100% of the time, the authors assigned greater weightings to top experts i.e. experts who historically posted more accurate predictions than other experts. To do this, they adopted a "mixture of experts framework", which "observes the predictions of a group of experts and combines the individual predictions into a single prediction... [and] automatically learns which experts are typically most accurate" (*ibid*, p.4). This framework differentiates this study from other message board studies we have already discussed such as Das and Chen (2001) and Antweiler and Frank (2004).

In addition to discussing individual stocks, experts often post optional sentiment tags, which indicate their prediction or recommendation i.e. 'strong buy', 'buy', 'hold', 'sell', or 'strong sell'. For that reason, Liu et al did not need to focus their efforts on extracting sentiment from the postings; they felt that the sentiment tags were

sufficient indicators. They used 71 datasets of messages with each dataset containing messages about a different stock symbol. The dataset spanned a period of just under twelve months. It is important to remember at this stage that experts sometimes have hidden agendas and do not always post genuine predictions; the authors referred to the strategy of "pump[ing] and dump[ing]" (*ibid*, p. 13), whereby experts post numerous 'strong buy' and 'buy' indicators to cause other posters to purchase large quantities of the stock and then they sell the stock at a quick profit.

In the "mixture of experts framework", multiple predictions by the same expert were aggregated into a single prediction; likewise, multiple conflicting predictions were aggregated into a single prediction. They used a number of strategies when evaluating their framework—they experimented with "all awake experts" (they ignored "sleeping experts" or those experts who did not make any predictions at time  $t$ ), "top experts" (they used only those with the highest weights at time  $t$ ), "worst experts" (they used only those with the lowest weights at time  $t$ ), and "threshold experts" (they used only those with the highest weightings in the past). One limitation of the "all awake experts" is that they could potentially all have low weights and therefore might not be very reliable; the "top experts" and "threshold experts" were introduced to counteract that problem. They used the "worst experts" to examine the effects of using experts with low weightings. Their baseline strategy was the average of "all awake experts" (unweighted).

To evaluate the various trading strategies, they used the average rate of return as the to-be-predicted variable. The rate of return was adjusted for market effects using the price of the stock, the price of the S&P 500 index, and the beta-value of the stock. The beta-value of a stock is "a measure of how much the stock fluctuates with regards to the market" (Liu et al 2006, p.7). Overall, they found that the "threshold experts" strategy outperformed all other strategies. They also found that the "top experts" and "all awake experts" performed roughly the same; on closer inspection they discovered that on some days, very few experts made recommendations and those experts also had higher-than-zero weightings (i.e. they were ranked as having some value). Therefore, the "all awake experts" also happened to be the "top experts". Unlike most of the studies we have discussed thus far, they also ran the same experiments taking

into account transaction costs. Once a gain, the "threshold experts" strategy outperformed the other strategies and a positive rate of return was only possible when weighted trading strategies were used.

The authors cited some limitations of their own approach and some areas where further work was needed. In particular, they said it would be ideal if one could ensure that the sentiment tags did not contain any misleading recommendations. They also said that it is not clear from the recommendations how long one should hold on to a stock after following up on a 'strong buy' or 'buy' recommendation. Finally, they could not say whether or not posters tend to go along with the consensus viewpoint of other posters and whether readers of the postings tend to follow posters who have proven to be correct in the past. Liu et al also cited some problems with earlier approaches, such as the approach taken by Das and Chen (2001) and Antweiler and Frank (2004), which assumed that all posters' contributions were accurate, that all contributions were equally important, and that all the posters contributed the same number of postings per day.

Lavrenko et al (2000) developed e-analyst, a system that used Bayesian language modelling (LM) and piecewise linear regression to align 38,469 Biz Yahoo! news stories relating to 127 stocks, with trends of time series. External relevance assignments from Yahoo were used to determine stories relevant to a particular stock. The system recommended news stories that were likely to have been derived from one of five trend types; the trend types were: a 'surge', 'slight rise', 'no recommendation', 'plunge', or 'slight fall'. To evaluate the performance of the system, they used a number of approaches.

Firstly, they used detection error tradeoff (DET) curves to evaluate the usefulness of the language models to predict trends. They found that they could achieve 10% recall (or 90% miss) with a false alarm rate of only 0.5%, which means that even though the user would only be alerted to 10% of the stories that would probably be followed by a 'surge', it is unlikely that many of those would be false alarm alerts. An increase in recall (e.g. to 40%) would lead to an increase in the false alarm rate (to 15%). As a baseline IR approach, they also compared the false alarm probabilities of their LM

approach with the false alarm probabilities of the more traditional vector-space (VS) approach and found that the LM greatly outperformed the VS for three of the four trend types, particularly at lower levels of recall, when predicting five hours ahead. Only 'plunges' yielded a similar result for both approaches. With the exception of these 'plunge' trends, they found that the errors made by the baseline VS approach were very close to those that would be made by random ranking.

Using a separate language model for each stock, the authors evaluated the profit-making potential of the system by implementing a basic trading strategy over a forty-day period: the investor bought when an upward trend was predicted and sold when a downward trend was predicted. Even though this simulation experiment only yielded a modest profit of \$280,000, the performance was found to be statistically significant at the 1% level.

They also found that simultaneous alignment of trends and news releases yielded much better performance than when they tried to align trends with documents that preceded the trend by one, five, or ten hours. Also, whilst stock-specific LMs yielded higher profits, they found that the profit varied significantly from company to company and companies that were rarely covered in the news were at a disadvantage due to their small training sets. Universal LMs, on the other hand, yielded lower overall profit but were deemed to be more stable as they trained one set of models for all stocks. As a result, they concluded that the best results might be achieved using a combination of stock-specific language models and universal language models (the latter could be used for smoothing).

Critics of the Lavrenko et al study include van Bunningen (2004) (see Section 3.3), who said that Lavrenko et al ignored temporal ordering. In other words, when they tried to predict the influence of a test article, it was very likely that another article that also caused the same trend was in the training set. Because any articles that caused the same trend probably had the same features, van Bunningen argues that this is not true prediction although we believe it is still prediction in the general sense. Kroha et al (2006) also argued that it is not possible to isolate market responses for one news article and that the same article can cause different reactions depending on investor

sentiment, current position, and other available news<sup>21</sup>; for that reason, they used a collection of news stories and they looked at long-term trends. Mittermayer and Knolmayer (2006b) also criticised the study. They disagreed with the assumption that it would have been possible to enter the stock market at the time of the news release and exit whenever a 1% or more profit was generated, because this latter exit strategy does not take stop loss transactions into account. As discussed earlier in this section, Mittermayer and Knolmayer (2006a) catered for stop loss transactions by defining a two-sided early exit strategy. They also criticised the choice of stocks because they argued that in real-world implementation, one would have to predict noisy stocks, as well as highly-volatile ones. Nonetheless, we believe that an examination of volatile stocks is a valid research pursuit in itself, as it implies that the market is frequently reacting to information. Other critiques include the fact that Lavrenko et al omitted transaction costs and they seemed to have unlimited and unrealistic funds during the simulation (Mittermayer and Knolmayer 2006a). Mittermayer and Knolmayer argue that even highly creditworthy investors who borrow money, typically only invest a single-digit multiple of the originally-available amount; however, in the Lavrenko et al study, they invested, on average, a multiple greater than 40. Despite these limitations, this study also has many strengths—several evaluation metrics were used, including random guessing, comparison with baseline approaches, prediction accuracy, detection error trade-off (DET) curves, and a trading strategy.

Koppel and Shtrimberg (2004) downloaded short news stories from the M ultex Significant Developments corpus<sup>22</sup> with a view to determining automatically the market reaction to the stories. For their feature set, they initially selected words that appeared 60 times or more. They then eliminated function words, with the exception of words such as *above*, *below*, *up*, and *down*, as they deemed these to be relevant. As the stories were relatively short—there were on average 100 words in each—they represented each story as a binary vector. Using their initial selection of features, they then considered only the 100 features with the highest information gain. With

---

<sup>21</sup> In Section 2.4, we also mentioned that market anomalies that have nothing to do with the arrival of information can impact on prices. Such anomalies are beyond the scope of this research however.

<sup>22</sup> Information about the M ultex corpus can be found on <http://aune.lpl.univ-aix.fr/projects/multext/MUL4.html>

regards matching stories to stock prices, they used two labelling approaches; the first involved matching each story with the change in the closing price the day before publication and the opening price the day after; the second involved matching each story with the change in the opening price the day after publication and the opening price the day after that. Stories were labelled as being positive if the stock price rose 10% or more and stories were labelled negative if the price decreased by 7.8% or more; the reason for the different thresholds was to ensure there were an equal number of positive and negative stories. Using these thresholds, their dataset comprised of 425 positive and 426 negative stories. Using a linear SVM, their first labelling approach, and 10-fold cross validation, they achieved an accuracy of 70.3%. When they trained on the 2000-2002 corpus and tested on the 2003 corpus, the algorithm yielded an accuracy of 65.9%. Boosting and selection of kernels did not have much impact on these results. They reported that naïve Bayes and decision tree learners yielded very similar results.

On closer inspection of the features, they identified certain features as being clear indicators of negative documents—these features included *shortfall*, *negative*, and *investigation*. They also reported that documents that contained those words, were nearly always negative and that the twenty words with the highest information gain in the corpus were all negative indicators. They found that recall for positive stories was high (83.3%) but precision was lower (66.0%) and that misclassification tended to occur mostly in negative documents which did not have any of the negative indicators.

When they used the second labelling approach and 10-fold cross validation to evaluate the potential profitability of the system, they only achieved an accuracy of just over 52%. Whilst they reported that this result bears out the efficient markets hypothesis (see Section 2.2 for a discussion), there are some limitations to their study that could disprove this, namely the size of the data set, the features chosen (they suggested that word collocation could prove helpful as co-occurring words could provide richer information), and their labelling approaches (they also stated that the use of prices immediately after publication of the story might prove helpful).

Kroha and Baeza-Yates (2004) used collections of online news stories to predict long-term rather than short-term trends, because they argued that it is not always possible to evaluate the impact of a news item at the time of its release. They manually approximated the trends and divided approximately 400,000 German news items from October 1999 to September 2003 into four trend sets (up1999, down2000, down2002, and up2003). Each trend set contained 16 documents which spanned 16 weeks and each set contained approximately 30,000 news items. They also collected the DAX30 index outcomes for the same period. Interestingly, they found that there were significantly more messages during the two up trends (32,299 and 35,998), than there were during the two down trends (30,228 and 23,875). They also found that there was a high number of unique words during all four trends. It should be pointed out at this stage that the system used to identify unique words (called *Bow*), counted hyphenated words as multiple words. In particular, they found that there were more unique words in the two up trends (212,314 and 240,151) than there were in the two down trends (111,249 and 150,878), which implies that the authors used a rich and varied vocabulary set, particularly during the up trends.

Their first hypothesis stated that positive words are more probable during positive (up) trends and negative words are more probable during negative (down) trends. In the first set of experiments, they used the Inverse Document Frequency (IDF) of substrings for manually-chosen keywords<sup>23</sup>. Whilst the IDF proved logical for some substrings when using a basic word set (e.g. see *gewinn* in Table 3.4, which appeared more in up trends than down trends), it was not logical for other substrings (e.g. see *negative* in Table 3.4, which appeared more in up trends than down trends). Overall, they reported that the results were only 50% valid (logical), so they did not prove this hypothesis.

<i>Substring</i>	<i>Up1999</i>	<i>Down2000</i>	<i>Down2002</i>	<i>Up2003</i>	<i>Valid?</i>
<i>Gewinn (profit)</i>	<b>27.96%</b>	24.39%	18.34%	<b>29.79%</b>	Yes
<i>Negativ (negative)</i>	<b>4.11%</b>	4.10%	5.22%	<b>7.59%</b>	No

Table 3.4: Inverse document frequency of one positive and one negative substring in the basic word set (Kroha and Baeza-Yates 2004).

<sup>23</sup> It is not clear how they devised the list of keywords.

In the second set of experiments, they used only the first 1,000 words with the highest probabilities for two of the trend sets (up1999 and down2000) and subsequently filtered them to find positive and negative substrings. They found that negative substrings tended to appear more during down trends than up trends, but they could not prove the positive substrings aspect of the hypothesis. Table 3.5 shows the inverse document frequencies for one positive and one negative substring with high probabilities.

<i>Substring</i>	<i>Up1999</i>	<i>Down2000</i>	<i>Valid?</i>
<i>Gewinn (profit)</i>	22.69	<b>24.39</b>	No
<i>Fallen (to fall)</i>	6.98	<b>8.08</b>	Yes

Table 3.5: Inverse document frequency of one positive and one negative substring with high probabilities (Kroha and Baeza-Yates 2004).

To prepare for Naïve Bayes classification, they randomly selected half of the documents for training (8 documents each for up1999, down2002, and up2003 and 9 documents for down2002) and used the other half for testing in 25 trials. Using the basic set of substrings, they achieved an average classification accuracy of 94.44%. However, when they performed classification using the probabilistic indexing method (i.e. with the most probable substrings, they achieved an average accuracy of 71.3%.

Kroha and Baeza-Yates then used the messages in all four trend sets for training and devised a new testing class (now2003) comprising 8 documents with 8 weeks of news items. The 8 documents of the class now2003 were classified as up2003, which they deemed to be an accurate reflection as the up2003 trend continued into the current trend (now2003). However, when they used the probability indexing method for classification, the 8 documents were classified as down2002. They then tested with unseen up/down trends to see if they would be classified like the up/down trends in the training collection but they did not achieve satisfactory results. The unseen up1999 trend was classified as down2000, down2000 was classified as up1999, down2002 was classified as up2003, and up2003 was classified as down2002. They suggested that unseen items tended to be assigned the next closest trend in terms of time period rather than in terms of features.

In their final set of naïve Bayes experiments, they appended the two up trend sets (32 documents) and the two down trend documents (32) to create up and down trend documents. They also used the now 2003 document. Using 50% of the data for training and the remainder for testing, they achieved an average accuracy of 75.69% over 8 trials, with a three-way classification. Using the same training and testing set with the probabilistic indexing method, they reported an average accuracy of 50%. In this study, evaluation of their results was limited to prediction accuracy.

In a related follow-on paper, Kroha et al (2006) also proposed that statistically, during growing markets, the contents of business news items should be different than the contents of business news during falling markets. They assumed that if this hypothesis was true, there should be more positive words during a growing market than negative words, and vice-versa. They also assumed that the probabilistic profile of news during growing/falling markets should be useful for classification purposes. They collected over 400,000 news items from October 1999 to November 2005 (a longer period than the previous paper) and the DAX30 index outcomes for the same period. Like in the previous paper, they manually approximated the trends and identified a down trend in 2000 and an uptrend in 2003. They then divided the news into an up and down trend set and ignored any news items that appeared when there was no clear trend (see Table 3.6 for details of each trend set).

<i>Trend set</i>	<i>No. of Weeks</i>	<i>No. of Training News Items</i>	<i>No. of Testing News Items</i>
<i>Up</i>	139	200,406	77,392
<i>Down</i>	132	186,146	34,678

Table 3.6: Breakdown of news items for each trend set (Kroha et al 2006).

Unlike the previous paper, they introduced a data cleansing stage prior to classification, whereby they removed duplicate news items, items that had been published in English, and non-relevant text such as the names of authors and towns. Using their own list of positive and negative words, they then counted the number of positive and negative words and found that positive words were in the majority even during the down trend, therefore disproving their first assumption.

With regard to the second assumption, they decided to group all the down trend news items together and all the up trend news items together, rather than classify each news item separately, mainly because they said it would be too time-consuming. For each week, they used 87% of the news for training with naïve Bayes and 13% for testing purposes. Whilst their reporting of results was limited, they did report an accuracy of about 75% in each class and a drop in accuracy to about 55% before the trend changed. They proposed that this latter result might be improved if each news item was labelled individually. Whilst they concluded that more research was needed, they also suggested that the delay between the change in positive/negative news and the market trend change might be explained by market psychology. They suggested that during a growing trend, investors are optimistic and fail to react quickly to negative news when it arises. Likewise, during a falling trend, investors are pessimistic and fail to react quickly to positive news. Such market behaviour would appear to contradict the efficient markets hypothesis. Like the previous study, evaluation metrics were limited to prediction accuracy; trading strategies, random guessing, and comparison with baseline approaches were not employed.

Tetlock (2007) examined the influence of the Wall Street Journal's (WSJ) 'Abreast of the Market' column, on stock market returns. The WSJ column recounts market activities from the previous day. Tetlock's column data spanned a 16-year period from 1984 to 1999. The main focus of his research was media pessimism, as he held the assumption that high media pessimism is correlated with low investor sentiment and this will result in downward pressure on prices. Using the General Inquirer (GI) content analysis program, he counted the number of words in 77 GI categories<sup>24</sup> and used the maximum variance in these categories to devise a single media factor. Because the factor was strongly related to pessimistic words, he referred to it as a pessimism factor. He used the Wharton Research Data Services (WRDS) to access nearly 4,000 time series observations of returns for the same time period.

---

<sup>24</sup> The General Inquirer program uses categories from the 'Harvard' and 'Lasswell' dictionaries to count the number of words in each category. Words can belong to more than one category and some words do not belong to any category. The Tetlock (2007) study used categories from the 'Harvard-IV-4' dictionary.

He performed various regression analyses using this pessimism factor and the 'negative' and 'weak' GI categories. His regressions produced many interesting results but only the most relevant results are outlined here. He found that the pessimism factor had a significant negative impact on the next day's returns ( $p$ -value  $< 0.001$ ) but the impact was temporary and fully reversed within a week. He also found that unusually high or low levels of pessimism led to temporarily high levels of trading volume. Finally, he found that pessimism had a particularly high impact on price for small stocks and that the impact tended to be slow to reverse itself. To evaluate the performance of the system, he constructed a hypothetical trading strategy using negative words and reported excess returns of 7.3% per annum. However, as this was a zero-transaction cost strategy, and inflation and spread costs were also not taken into account, it is unclear whether or not it would really be worthwhile.

In a related paper, Tetlock et al (2008) also used the 'negative' GI word category and regression analyses to quantitatively measure language and predict individual firms' earnings and returns. They proposed that if the other two sources of fundamental information, namely analysts' forecasts and accounting variables, are biased or incomplete, then linguistic content may have some additional explanatory power for explaining future earnings and returns. They examined the impact of negative words in all WSJ and Dow Jones News Service (DJNS) stories for S&P 500 firms from 1980 to 2004.

They found that negative words succeeded in conveying negative information about earnings above and beyond that conveyed by analysts' forecasts and historical data. For example, the coefficient estimates of the ability of negative words to predict lower standardised unexpected quarterly earnings (SUE) using the DJNS and WSJ sources were  $-4.42$  and  $-5.28$  respectively. They also found that the reaction to negative information usually takes place within one day and proposed that profits could be made if one used news sources that are updated frequently throughout the day (e.g. the DJNS); nonetheless, one would have to carefully consider how any such profits could be diminished by transaction, inflation, and other related costs. Even though they identified a clear reaction to negative news, they deemed the market to be relatively efficient as the reaction was prompt and the returns were relatively small in

the days that followed release of the news. They found that negative words in stories which discuss fundamentals (such as cash flow or return on investment) predicted earnings more effectively than negative words in stories about other issues or events. In other experiments, they examined the impact of reasonable transaction costs on the profitability of their trading strategy and found that it was no longer profitable. However, they could not rule out that more sophisticated trading rules might minimise the impact of transaction costs.

Engelberg (2008) examined the transaction costs associated with processing soft qualitative information compared to hard quantitative information. He suggested that because the processing costs of complex qualitative information can be high, it is possible that some information will not be incorporated into prices immediately and that they may take some time to do so.

Like Tetlock (2007) and Tetlock et al (2008), Engelberg used the Harvard Psychological Dictionary via the General Inquirer (GI) program to count the number of negative words, thereby measuring the soft qualitative content. However, unlike the previously-mentioned authors, Engelberg focused specifically on DJNS articles about earnings announcements, as they tend to contain both hard and soft information and are repeated over time. See Table 3.7 for an overview of the characteristics of hard and soft content of earnings news, as defined by Engelberg. Using five different sources for the hard and soft information, he collected 51,207 earnings announcements between January 1999 and November 2005 relating to 4,700 firms. One limitation of his study, as identified by Engelberg, was a bias against small firms, as larger firms tend to feature more in DJNS articles. Like Tetlock et al, he measured the hard quantitative content of the earnings 'surprise' using the standardised unexpected earnings (SUE), which he defined as the difference between the median of the analysts' forecasts and the firm's actual earnings per share, scaled by a normalisation factor.

	<i>Hard earnings news</i>	<i>Soft earnings news</i>
<i>Based on:</i>	Accounting data (earnings)	Text of media articles written about earnings
<i>Qualitative or quantitative?</i>	Quantitative	Qualitative
<i>Easily comparable across firms?</i>	Yes	No
<i>Independent of collector?</i>	Yes	No (not everyone may interpret it the same)
<i>Easy to store?</i>	Yes	No
<i>Easily passed on without loss of information?</i>	Yes	No

Table 3.7: Definition of hard and soft content of earnings news (Engelberg 2008).

Engelberg downloaded the headline and lead paragraph of each article, on the assumption that the DJNS journalists summarise the content in the headline and lead sentence. Using the GI program, he then calculated the fraction of negative words as follows:

$$\text{Negative Fraction}_{it} = \frac{\text{total number of negative words for firm } i \text{ on day } t}{\text{total number of words for firm } i \text{ on day } t}$$

Using this somewhat crude measure, he found that almost half (47.5%) of the articles had no negative words in the headline or lead paragraph, implying that the remainder (52.5%) did have negative words.

Before examining the relative predictability of soft versus hard information, he investigated the predictable capabilities of soft information beyond those of the SUE. He sorted the articles into SUE quintiles (fifths) using the previous period's calendar quarter to determine the SUE cut-offs. He found that the post-earnings announcement drift (PEAD) increased across the SUE quintiles. The lowest quintile experienced an average 80-day cumulative abnormal return (CAR) of -0.60% and the highest experienced an average 80-day CAR of 1.99%. He found the difference of 2.59% to be statistically significant using both a parametric and a non-parametric test. He then sorted each quintile into one of three 'bins' based on their negative fraction, as calculated above. The first bin comprised articles with a negative fraction of 0% (no negative words), the second comprised articles with negative fractions greater than 0 and less than 5%, and the third comprised the remaining articles (negative

fraction >5%). Like Tetlock et al (2008) he found that the qualitative information had additional predictable capabilities beyond those of the SUE and that this information was not immediately incorporated into prices. For example, within SUE quintile 5, the average 80-day CAR was 3.90% for firms with a negative fraction=0% and 0.28% for firms with a negative fraction > 5%. Engelberg suggested that negative word articles for firms in low SUE bins might simply be reiterating the hard quantitative information and therefore the negative words might not contain high predictive capabilities. On the other hand, he suggested that articles for firms in high SUE bins which contained negative words might contain additional information that had predictive uses. When he performed regression analyses using five different trading strategies, three of which incorporated soft earnings information, he found that all five generated statistically and economically significant profits at some stage.

As mentioned previously, Engelberg used negative fraction as a proxy for soft information and SUE as a proxy for hard information. He assumed that if soft information is more costly to process, it should diffuse more slowly into prices relative to hard information. Additional regression analyses found that negative fraction was the only statistically significant predictor of the next earnings announcement return and it continued to be a predictor for the next two, three, and four quarters. He interpreted this finding as evidence that negative fraction is more costly to process than SUE and CAR[-1,1] which is why it diffuses more slowly into prices.

In another experiment, he used the Stanford parser to perform typed dependency parsing<sup>25</sup>, to determine what kinds of soft information predict returns. He devised six categories of earnings news: 'positive fundamentals', 'negative fundamentals', 'future outlook', 'environment', 'operations', and 'other'. 'Positive fundamentals' included *earnings*, *sales*, and *revenue*; 'negative fundamentals' included *costs*, *expenses*, and *charges*. Words associated with 'future outlook' included *outlook*, *plans*, and *forecast* and words associated with 'environment' included *conditions*, *customers*, and

---

<sup>25</sup> Typed dependencies are also known as grammatical relations. The Stanford parser, which is available for download on <http://nlp.stanford.edu/software/lex-parser.shtml>, can assign 48 grammatical relations to text.

*economy*. 'Operations' related to words such as *business*, *production*, and *services* and 'other' covered all words not in any of the other five categories. He then calculated the fraction of typed dependency pairs between negative words and each of these six categories. When he replaced negative fraction with the six new variables, and performed regression analyses to predict future returns, he found that the coefficient on each of the six variables was negative, with the exception of negative fundamentals. Engelberg suggested that this result made sense, as negative words used with negative fundamentals can have a positive meaning (e.g. 'low costs') or negative meaning (e.g. 'disappointing costs'). Of the other five categories, he found that costly soft information about 'positive fundamentals', 'future performance', and 'other' were the only categories that had statistically and economically significant negative coefficients and therefore these categories are the most important for predicting future returns. He also suggested that the 'other' category needed to be explored further as it was a significant category.

For robustness, Engelberg split his sample into two periods and re-ran his regressions, to take into account two key events which could have impacted his results—the bursting of the Internet Bubble in March 2000 and the implementation of the Regulation Fair Disclosure in October 2000. He found some support for the hypothesis that soft information was a better predictor before the bubble burst and hard information was better afterwards but he still reported that soft information diffused more slowly into prices during each period. Some strengths of this study include his use of statistical techniques and the addition of additional variables (events) to measure robustness and his adoption of a trading strategy to evaluate profitability. However, they did not take transaction or other related costs into account.

Henry (2008) examined the market impact of tone and other style aspects in 1,366 earnings press releases relating to 562 firms. The other style aspects examined were length, numerical intensity, and verbal complexity. The firms were all related to the telecommunications, computer services, or related equipment manufacturing industries.

She used the collocation feature of Wordsmith Tools to decipher whether directional words such as *increased* and *greatest* were inherently positive or negative, depending on the context of their occurrences in the press releases. She defined positive (negative) words as those which appeared near desirable (undesirable) financial items. Examples of desirable financial items included *dividends* and *cash flows*; undesirable items included *losses* and *expenses*. She used the Diction software to calculate the frequency of positive and negative words in the press releases, thereby measuring the tone, and then used regressions to correlate positive (negative) tone with abnormal positive (negative) returns. The mean value of tone was 0.568, which suggests a bias towards positive tone in the press releases<sup>26</sup>. She also found that abnormal returns tended to be higher when the tone was positive (the correlation between abnormal returns and tone was 0.098). However, with regard to the other variables, she found that lengthier press releases, and to a lesser extent, releases with high levels of numerical intensity, tended to diminish the positive impact of unexpected earnings on returns. She did not find any correlation between verbal complexity and impact on returns.

Henry cited some weaknesses of her own study which included an inability to capture the subtleties and complexities that may be hidden in individual firms' press releases. She also suggested that measures such as the Flesch and Fog indices, the latter of which was subsequently critiqued by Loughran and McDonald (2011a) (see Section 2.3), could have been used instead to measure verbal complexity. In addition, a broader selection of industries might have yielded different results; we attempt to cater for this in our own research study by not being industry-specific. Another limitation of this study is that she used limited evaluation metrics (prediction accuracy).

Loughran and McDonald (2011b) devised a list of 2,337 financial negative words (Fin-Neg) with a view to comparing the performance of the list with the Harvard-IV-4 TagNeg (H4N) negative word list. They argued that the H4N list contains many words that are not negative in a financial sense e.g. *tax*, *cost*, *liability*, and *crude* (oil).

---

<sup>26</sup>A value of -1.0 would indicate a completely negative tone; +1.0 would indicate a completely positive tone.

They also proposed five additional words lists (positive, uncertainty, litigious, strong modal, and weak modal) which could be used in future research. They decided to create these exhaustive lists<sup>27</sup> rather than “let the data determine the most impactful words” (p.44) as these exhaustive lists would make it more difficult for managers to select specific words to avoid.

To create the words lists, they generated a dictionary of words and word counts from all the 10-Ks. They then manually examined all the words that occurred in at least 5% of the documents and considered their likely usage in financial documents. They only considered simple negation for the Fin-Pos (positive) words as they did not expect to see simple negation for negative words; an example of the latter would be “not terrible earnings”, which is unlikely to appear in a document (*ibid*, p. 44). However, one flaw of this approach is that avoiding simple negation for negative words will not catch instances such as “losses lower than expected”.

In addition to using simple proportion of words (based on word count), they used the TF\*IDF term weighting scheme with the vector space model (aka bag-of-words), to ensure that common words would be weighted less. Whilst they found that the weighting scheme minimised noise, they found that both the H4N and FinNeg lists produced similar results.

Even though there were more negative words in the Management Discussion and Analysis (MD&A) section of the 10-K, than in the full 10-K, there were less Fin-Neg words than H4N-Inf words. They also found that there were “considerable” misclassifications of negative words in 10-Ks when using the H4N list (p.47). For example, commonly-used business words such as *tax*, *costs*, *loss*, *capital*, *cost*, *expense*, and *expenses*, which are deemed negative in the H4N, accounted for more than 25% of the negative words—yet, these are not necessarily negative words in a financial context. They argued that some of the Harvard negative words may in fact be acting as a proxy for other industry effects, and therefore proposed that their customised Fin-Neg wordlist was a better indicator of negative news. Overall, they found that the MD&A section does not appear to have a greater impact on excess

---

<sup>27</sup> See [http://www.nd.edu/~mcdonald/Word\\_Lists.html](http://www.nd.edu/~mcdonald/Word_Lists.html) for each of the word lists.

returns than the entire 10-K, although they did point out that it was a smaller sample size.

Using regressions with and without multiple control variables (e.g. firm size, book-to-market, share turnover, and institutional ownership) Loughran and McDonald examined the impact of their Fin-Neg list on excess returns. They also controlled for cross-sectional effects on the data. They defined excess return as “the firm’s common stock buy-and-hold stock return minus the CRSP value-weighted market index buy-and-hold return” (p.51). They used a four-day window from days 0 through day 3, where day 0 is the 10-K filing date.

When they broke firms into quintiles based on proportion of Fin-Neg words, they found a strong pattern between the number of negative words and decreased returns. When they tried to link negative word lists with filing period returns, they found that firms in quintiles where there was a lower percentage of negative words had slightly negative returns in the four-day window but firms in quintiles with a high percentage of negative words had greater negative returns. Even when they performed regression analyses controlling for other variables, they found that “only a small amount of the variation in filing period returns is explained by the independent variables” (p.53).

When they calculated the portfolio returns generated by taking a long position on stocks with a low negative word count and a short position on those with a high negative word count, they found that negative tone was not related to future returns if one was considering a trading strategy.

They also considered two sub-samples of 10-Ks: (1) 10-Ks filed by companies accused of accounting fraud and therefore subject to shareholder litigation and (2) 10-Ks filed by firms disclosing at least one material weakness in internal control. Whilst they did not know whether to expect more or less negative words in such disclosures (companies sometimes disguise negative problems), they found that firms with a higher proportion of negative financial words or strong modal words were more likely to report material weaknesses. When they introduced term weighting, they found that

firms that used stronger language (more positive, more negative, or more modal strong words) were more likely to disclose material weaknesses. Both these findings suggest that word lists are helpful features in the analysis of 10-Ks. Some strengths of this study include their use of various evaluation metrics (simple word count, prediction accuracy, and adoption of a trading strategy), and their use of multiple control variables.

In this section, we examined a number of studies that used single words as features, to analyse and classify financial documents. Looking at Table 3.8, we can see that none of the studies re-used word lists from other studies, although some studies used similar methods to devise their word lists. For example, Antweiler and Frank (2004), Gidofalvi (2001), and Koppel and Shtrimerberg (2004) used the top 1,000 words as ranked by information gain (although they used different data sources). In addition, Tetlock (2007), Tetlock et al (2008), and Engelberg (2008) all used the negative General Inquirer (GI) category in their studies. The range of features used suggests that researchers believe that the features need to be customised to the data source being examined (e.g. press releases, forum postings, and disclosures).

Das and Chen (2001)	Used a hand-selected lexicon of financial words statistically processed to determine strongest discriminators.
Swales and Yoon (1992)	Performed content analysis of similar words and phrases alluding to themes.
Schumakar and Chen (2006)	Used three representations to represent documents (bag-of-words, noun phrases, and named entities).
Mittermayer and Knolmayer (2006a)	Used a hand-made thesaurus of words and phrases that had an impact on prices (single words, phrases, and tuples of words and phrases). Used the thesaurus with a bag-of-words model.
Fung et al (2005)	Used the vector space model to represent each document.
Antweiler and Frank (2004)	Used the top 1,000 words, as ranked by information gain.

Antweiler and Frank (2006)	Identified news topics. Used the top 1,000 words, as ranked by information gain.
Gidofalvi (2001)	Used the top 1,000 words, as ranked by information gain.
Liu et al (2006)	Used sentiment tags posted by experts.
Lavrenko et al (2000)	Used language models (stock-specific and universal).
Koppel and Shtrimberg (2004)	Used words appearing 60+ times (excl. most function words). Then considered top 1,000 features, as ranked by information gain.
Kroha and Baeza-Yates (2004)	Used IDF of substrings for manually-chosen keywords. Also experimented with the top 1000 words with the highest probabilities.
Kroha et al (2006)	Used own list of positive and negative words.
Tetlock (2007)	Counted the number of words in 77 General Inquirer (GI) categories. The GI program uses categories from the Harvard and Lasswell dictionaries to count the number of words in each category. Focused on the negative and weak GI categories.
Tetlock et al (2008)	Used the negative GI word category.
Engelberg (2008)	Used the negative GI word category and calculated the fraction of negative words. Also devised categories of earnings news, comprising words.
Henry (2008)	Used the collocation feature of WordSmith Tools to examine if directional words were inherently positive or negative. Defined positive (negative) words as those which appeared near desirable (undesirable) financial items. Used the Diction software to calculate the frequency of positive and negative words.
Loughran and McDonald	Generated a dictionary of all words with their word

(2011b)	<p>counts.</p> <p>Then examined all the words that occurred in at least 5% of the documents and considered their likely financial usage. Used these to devise a list of financial negative words. Also devised lists for positive, uncertain, litigious, strong modal, and weak modal.</p> <p>Also used TF*IDF to ensure that common words would be weighted less.</p>
---------	--

Table 3.8: Methods used to generate word lists (various studies).

### 3.4 Keyword Records and Phrases

Peramunetilleke and Wong (2002) analysed the content of online real-time macroeconomic news. However, instead of predicting the future performance of companies (our research goal), they predicted hourly/intra-day foreign exchange rates, using a rule generation algorithm. The results were then compared with those from regression analysis of time series data and with those from a neural network. As input, they used news headlines and over 400 keyword records chosen by a domain expert. Unlike other studies that used single keywords (see Section 3.2), their keyword records consisted of sequences of words—word pairs, triples, quadruples, and quintuples. The complete listing of keywords was not published in the paper; only some sample keyword records were provided, such as *US inflation weak* and *pound lower*.

Even though they experimented with different time periods (one, two, or three hours), with different weighting methods (the Boolean method, the TF\*IDF method, and the TF\*CDF method), and with different currencies (DM/USD and JPY/USD), they used the same headlines and keyword records in each experiment. The TF\*CDF method is an alternative method for calculating the weight of a keyword record. The term frequency (TF) is multiplied by the category discrimination frequency (CDF), instead of the inverse document frequency (IDF). The CDF of keyword A is equal to the sum of its category frequencies ('up', 'down', and 'steady') divided by the number of time

windows containing at least one instance of A. The weight is then normalised.

Weighted keywords and the closing values of the exchange rate (the last 60 time periods for which the outcome was known) were used to generate the classification rules; the accuracy of these rules was tested in previous experiments using weights that were manually assigned by foreign exchange dealers. In the manual assignment, dealers assigned weights to twelve economic factors, including the employment rate, inflation, and political news. Whilst the authors reported that the dealers found the rules to be “successful” (*ibid*, p. 136), unfortunately they did not provide specific results in this paper. They used three classifications—'up', 'down', and 'steady'—to refer to the direction of the foreign exchange rate. The size of the dataset was not stated.

The best automatic classification result (51%) was achieved using the TF\*CDF weighting method and the rule generation algorithm for a one hour time period for DM/USD. This result was first evaluated by comparing it to the best result from a statistical time series analysis (37%), using the same test and training period. Unfortunately, foreign exchange traders were unwilling to measure their own prediction abilities but they agreed that it would be difficult to achieve 50% accuracy. In another evaluation, using the neural network, the average accuracy was 37.5% using the last 2 to 10 outcomes (the outcomes were 'up', 'down', and 'steady'); this never reached 50%. They also found that DM/USD had a higher prediction accuracy than JPY/USD for each weighting method and for each time period.

When they compared their best result (51%) to random guessing (33%), the probability that the prediction accuracy would equal or exceed 51% when using random guessing and the TD\*CDF weighting method was found to be less than 0.4%, in 60 trials. They also found that the probability of achieving the average accuracy (48.6%) for DM/USD for all three time periods when using random guessing and TF\*CDF was less than 0.001%. Finally, the probability of achieving the average accuracy (44.3%) for JPY/USD for all three time periods when using random guessing and TF\*CDF was also less than 0.001%. Therefore, their system was deemed to be better than chance.

For DM and the best performing weighting method (TF\*IDF), the authors also provided the breakdown of percent totally wrong, percent slightly wrong, and percent accurate. It is important to note here that a 'steady' classification that should have been an 'up' or 'down' classification, was only considered 'slightly wrong'; the authors state this is a flaw in their study. The prediction accuracy of 'ups' and 'downs' for two hours for DM/USD showed the greatest inaccuracy:

- Their system predicted that 35% would go up but 30% were in fact up.
- Their system predicted that 22.6% would go down but 38.3% were in fact down.

It is not clear in the paper whether these 'up' and 'down' judgements were accurate judgements or not; it is only clear that the system *almost* judged the correct quantity of judgements as up or down. Kroha and Baeza-Yates (2004) disagreed with their method of classifying news so soon after its release, partly because different experts will react differently depending on the context of the news and also depending on overall market sentiment. The Peramunetilleke and Wong (2002) study differs from ours in many ways, in terms of the prediction goal (exchange rate fluctuations vs. future performance of companies) and the types of data used (news headlines vs. corporate 8-K disclosures). Nonetheless, their study is somewhat relevant to ours as they also used keywords for classification. Whilst they did compare their results to random guessing, they did not compare it to baseline approaches or evaluate it using a trading strategy (Mittermayer and Knolmayer 2006b).

Wüthrich et al (1998) developed a prediction system for five market indices—the Dow Jones Industrial Average (DOW), the Nikkei 225 (NKY), the Financial Times 100 (FTSE), the Hang Seng (HSI), and the Singapore Straits (STI). Their system counted the number of occurrences of certain sequences of keywords describing macroeconomic and microeconomic events in daily online Wall Street Journal news articles. 'Up', 'down', or 'steady' predictions were made available real-time before the first (Asian) markets commenced trading. They categorised an 'up' prediction as an increase of at least 0.5%, a 'down' prediction as a decrease of at least 0.5%, and a 'steady' prediction as a change of less than 0.5%. The system used an 'up', 'down', or 'steady' prediction with the latest closing value to predict the expected closing value.

They used over 400 keyword records chosen by a domain expert; unfortunately, only a sample of keywords was provided in the paper. Like Peramunetilleke and Wong (2002), their keyword records consisted of sequences of words (word pairs, triples, quadruples, and so on); single words were not used. Stemming algorithms were applied to ensure that different forms of the same keyword record could be found (e.g. the word pair *stock drop* would also match the sentence "stocks have really dropped").

The authors experimented with a rule-based algorithm, the  $k$ -NN algorithm, regression analysis, and a neural network. It is not clear, however, whether the following prediction results relate to the expected actual closing value or to the change in direction as the paper implies that there were different prediction goals for the different algorithms.

Using the rule-based algorithm, the best prediction results were achieved with the DOW, which was accurate 45% of the time and the FTSE, which was accurate 46.7% of the time. These two indices were 'slightly wrong' 46.7% and 36.7% of the time respectively and 'totally wrong' 8.3% and 16.6% respectively. The previous 100 stock trading days were used as training data to predict the following day's prices and 60 trading days were used for testing purposes. Using the  $k$ -NN algorithm, they achieved the best prediction results with  $k=9$  and the Euclidean similarity measure (FTSE 42%, NKY 47%, DOW 40%, HSI 53%, and STI 40%). Whilst they stated that the test period was shorter, they did not state how short. Each of these results compares favourably with the results that would be achieved with random guessing (33%). Using regression analysis on a 20-day moving average, they did not achieve 40% accuracy for any index. Using a neural network and the actual direction for the last  $n$  days (with  $n$  between 4 and 10), the average prediction accuracy was as follows: HSI 43.9%, FTSE 35.4%, DOW 36.8%, NKY 34.1% and STI 32.5%. Note how only the HSI achieved higher than 40%. In this experiment, 60 days were used for training and 40 days for testing purposes.

When they evaluated the performance of their trading strategy, the authors reported that they would have generated 7.5% profit on the DOW over three months, a s

opposed to the actual appreciation of 5.1%. The results for other indices were as follows: the FTSE yielded 5.5% (vs. 11% actual appreciation), NKY yielded 5% (vs. 4.3% actual appreciation), HSI yielded 3.5% (vs. -4.6% actual depreciation), and STI yielded 4.5% (vs. -8.8% actual depreciation). Like the simulations carried out by Lavrenko et al (2000) (see Section 3.3), these profits assumed zero transaction costs and that the investors would have been in a position to invest the same amount of money each day; in reality, the situation would be different. They also did not consider inflation or costs associated with the spread. In a critique of the study, Fung et al (2005) proposed that such a system should be able to predict intra-day prices as well as closing/opening prices. On the issue of closing/opening prices, Mittermayer and Knolmayer (2006b) said that the Wüthrich et al (1998) study assumed that the predicted closing price for one day would automatically be the same as the opening price for the following day. However, because the latter system is better at prediction than random traders, it will buy and sell differently to random traders and therefore the predicted value cannot reflect reality.

Thomas and Sycara (2000) experimented with four methods: (1) maximum entropy text classification (METC) using text from online stock discussion boards; (2) a genetic algorithm (GA) using rules based only on numerical data; (3) maximum entropy text classification and a genetic algorithm; and (4) maximum entropy text classification and multiple runs of a genetic algorithm. Whilst each of these methods were used for following-day closing price prediction, the authors were more interested in the profitability of the prediction, and so they only reported the excess returns that could have been generated, not the accuracy of the predictions.

They used 52 days for training and 252 days for testing purposes and they aggregated daily posts for 22 stocks to predict the best trading strategy. Unfortunately, they did not state how many posts were used in the experiments. Like Das and Chen (2001) and Antweiler and Frank (2004) (see Section 3.2), the Thomas and Sycara approach assumed that all posters' contributions were accurate, that all contributions were equally important, and that all the posters contributed the same number of postings per day. In reality, this would be highly unlikely. The numerical data used for method (2) comprised of trading volume, number of messages posted per day, and the

total number of words posted per day.

It is important to note at this stage that a probability of 0.5 was used for method (1); if the probability was greater than 0.5, it was considered an 'up' prediction and they recommended a 'buy-hold' strategy; otherwise, they recommended a 'sell' strategy. For method (2), they examined the percentage of the population that issued a 'buy-hold' signal.

Methods (1) and (2) yielded negative excess returns for all 22 stocks. However, when they only used the 12 stocks which had more than 10,000 posts during the year, small positive excess returns were achieved—6.91% for method (1) and 5.95% for method (2). However, when they used method (3) by integrating the METC and GA predictors, they reported excess returns of 2.9% for all 22 stocks and 19.3% for the 12 stocks which had more than 10,000 posts during the year.

When the authors used method (4) and averaged the prediction of multiple genetic algorithm runs and combined them with the METC, they achieved positive excess returns in excess of 30% for the same 12 stocks. Note how this return is significantly higher than the 19.3% achieved for method (3).

Unfortunately, they did not report on the accuracy of any of the predictions and they did not state whether or not 30% is indeed a satisfactory return. They only reported that method (3) yielded statistically significant results for the 12 stocks that had more than 10,000 posts during the year. Also, as pointed out in Schumaker and Chen (2006) (see Section 3.2), online postings are susceptible to bias and noise, so they are not the most reliable source of facts.

Several other authors have evaluated online news articles in an attempt to predict the future performance of companies. Cho et al (1999) used keyword record counting and novel weighting techniques for processing macroeconomic text in online news stories, before predicting the impact of the new stories on the daily movements of the Hang Seng Index (HSI). Domain experts devised 392 keyword records—examples include *dollar strong, not interest\_worry, dollar weak against mark, and interest rate*

*cut*—and each keyword record was allocated a weight for each class ('up', 'down', or 'steady') and each day. A change of 0.5% or more between today's closing value and yesterday's closing value of the HSI was classified as an 'up' or 'down'; changes less than 0.5% were classified as 'steady'.

The authors used forward source selection to select the most predictive individual sources and the most predictive combinations of sources from five web sites. The five web sites were the Wall Street Journal, the Financial Times, CNN, the International Herald Tribune, and Bloomberg. They experimented with four different weighting schemes and examined the maximum accuracy, the smoothness of the curve, and the source choices. Probabilistic rules were generated using the weights, and the rules were fed into a rule-based forecast engine, along with the closing values for the previous training days. 100 training days and 79 test days were used in the experiments. The four schemes were simple weighting (SW), vector weighting with class relevance (VWClassRel), vector weighting with class relevance and discrimination (VWClassRelDisc), and vector weighting with cluster relevance and discrimination (VWClusterRelDisc).

In the first set of experiments, where they did not use cross-validation and they combined between one and seven sources, the results were as follows (see Table 3.9):

<i>Weighting Schemes</i>	<i>Highest Accuracy</i>	<i>Comments</i>
<i>SW</i>	0.52 ("global max")	Using three sources; choppy curve, partly increasing, mostly decreasing.
<i>VWClassRel:</i>	0.475 (approx.)	Using two sources but these two were deemed insufficient on their own i.e. they would need to be supplemented; fairly smooth curve, mostly decreasing.
<i>VWClassRelDisc:</i>	0.475 (approx.)	Using six sources; fairly smooth curve, mostly increasing.
<i>VWClusterRelDisc:</i>	0.51	Using six sources which were deemed "intuitively reasonable"; smooth curve, mostly increasing.

Table 3.9: Results on test data without cross-validation and combining between one and seven sources (Cho et al 1999).

In the second set of experiments, where they used five-fold cross-validation and combined one to seven sources, the results were as follows (see Table 3.10):

<i>Weighting Schemes</i>	<i>Highest Accuracy</i>	<i>Comments</i>
SW	0.472	Choppy curve, partly increasing, partly decreasing.
VWClassRel:	0.43 (approx)	Choppy curve, partly increasing, partly decreasing.
VWClassRelDisc:	0.428 (approx)	Choppy curve, partly increasing, partly decreasing.
VWClusterRelDisc:	0.468 (close to global max of SW = 0.472)	Smooth curve, mostly increasing.

Table 3.10: Results on test data using five-fold cross validation and combining between one and seven sources (Cho et al 1999).

The authors found the VWClusterRelDisc scheme to be the best weighting scheme overall, even though it was also the most computationally demanding. Whilst the SW scheme achieved a slightly higher mean prediction accuracy (0.52 without cross-validation and 0.472 with cross-validation vs .0.51 and 0.468), the VWClusterRelDisc scheme had a smoother and mostly increasing curve. The probability that random guessing could achieve a high prediction accuracy of 0.468 was 0 and the probability that random guessing could achieve the worst prediction accuracy (0.402) was 0.0018.

When they examined the sources selected by VWClusterRelDisc when they did not employ cross-validation, they claimed that they were most similar to those that would be typically selected by a human expert. However, they only provided a brief discussion of the sources selected and they did not discuss the usefulness or reliability of the sources selected when they employed cross-validation. Some strengths of this study include their use of multiple evaluation metrics (random guessing and prediction accuracy) and their use of cross-validation and statistical techniques to ensure robustness.

Thomas (2003) manually created an ontology comprising 11 broad categories and 69 nested sub-categories, with each sub-category referring to a type of news headline. Examples of sub-categories included 'merger', 'shareholder meeting', 'SEC filing 8-K', 'lawsuit', and 'stock split'. He then built classifiers by hand for 39 of the 69 sub-categories with each classifier comprising a set of phrase detectors or regular expressions describing that category. The goal of this section of the study was to

classify news headlines into pre-defined categories. Thomas assumed that headlines could only belong to one category and that headlines contain sufficient information for classification; we do not agree that these assumptions will always hold true.

The dataset comprised one year of news headlines from Yahoo! Finance for all stocks in the Russell 3000. Yahoo! Finance contains headlines from numerous other sources, including EDGAR, Reuters, and Business Wire. He classified two weeks of headlines in March 2001 by hand; the first week was used to design rough classifiers and the second week was used to evaluate the precision and recall of the classifiers. The month of April was used to improve the classifiers, by examining false positives and false negatives. The month of May was used as holdout data, to evaluate the precision of the 39 classifiers. Due to the size of the dataset, and the time it would take to classify all the headlines for a full month by hand, he did not evaluate the recall for the month of May. The author found that precision was over 90% for 33 of the 39 sub-categories for the second hand-classified week and for 31 sub-categories for the full extra month. Recall was lower and much more inconsistent, particularly for categories which had few examples. For example, recall was under 30% for several categories which had less than 25 examples, and one category only had a recall of 6.3%.

He then used the classified news headlines in conjunction with a trading rule learner to signify stocks that should be avoided. He found that these results were better than when a simple 'buy-and-hold' strategy was used with a holdout dataset. As technical analysis is beyond the scope of our study, we will not discuss the development and performance of Thomas' trading rules in any detail here. Even though he found that the total amount of news performed better with the trading rule learner than individual categories of news, he proposed that with a larger dataset, individual categories might prove more useful than the total news. He also proposed that individual categories could prove useful for events studies (see Section 2.4 for a discussion). His preliminary results showed that some categories of news events with low relative occurrence (e.g. 'debt repayment' and 'price increases') resulted in positive returns (2.09% and 3.61% respectively) and other categories (e.g. 'earnings restatement' and 'options') resulted in negative returns (-2.29% and -6.24%

respectively). Unfortunately, to avoid duplicated events, he excluded any events that occurred within ten days of another event, so it is possible that some important events were ignored. Also, his classifiers were incomplete, as he only developed classifiers for 39 of the 69 sub-categories.

Thomas admitted that because the cumulative returns were derived from closing prices, any reactions to news that happened before the close of business, were included in the *previous* day's returns. Therefore, there could well be temporal issues in the data. Particular strengths of this study include his use of multiple evaluation metrics (classification accuracy, precision and recall, and a trading strategy) and also his consideration of transaction costs, liquidity, and spread issues.

Seo et al (2004) evaluated quantitative financial news articles from various sources including the CNN Financial Network, Forbes, Reuters, Motley Fool, and Business Wire but instead of trying to predict the future performance of companies, their goal was to summarise trends about the current financial performance of a company. Unlike many of the previous authors who also used online news articles, they only classified financial news articles i.e. those that referred specifically to the financial status of the company.

After eliminating stop words from 6,239 online news articles, they adopted the vector space model. Infrequent and high frequency words were eliminated and weights were assigned to the remaining words using a variant of the TF\*IDF method. Non-financial news articles, such as those referring to 'legal issues' or 'changes in corporate control', were filtered out using TextMiner, their information retrieval tool. Non-financial articles were still presented to the user via Warren, their user-agent system; they were just not used for classification purposes. The most informative co-located phrases were then identified using information gain (examples include *shares* and *fell* which could be selected from the sentence *Shares of ... fell by...*) and these phrases were used as experts in their "domain experts with vote entropy" classification system. This system relies on the vote of each expert to estimate the label for unlabelled data; each expert uses knowledge gained from the labelled training data. They compared the performance of this system with naïve Bayes with expectation-

maximisation, naïve Bayes with vote entropy, domain experts with expectation-maximisation, the most frequent class/category (note all unseen articles were labelled 'neutral'), and random guessing. The last two methods were used simply as baseline comparisons.

In the first experiment, they used a five-way classification ('good', 'good uncertain', 'neutral', 'bad uncertain', and 'bad') to determine how well the system used unlabelled data to improve classification accuracy. The two 'uncertain' classifications were introduced to allow for differences of opinion; for example, one human might label an article as 'neutral' whereas another human might label an article as 'bad' because it comes from an unreliable news provider, even though it may contain reasonably neutral content. They used 1,239 labelled articles for training and they carried out 50 iterations for each method; at each iteration, they used 50 unlabelled articles. Using this method, they made a "domain experts" classifier that was 75% accurate when they used 1,239 labelled articles and 450 unlabelled articles. The other five systems performed less well, although naïve Bayes with expectation-maximisation performed almost as well, achieving approximately 72% accuracy with the same quantity of labelled and unlabelled data. One of the reasons cited for the good result was the fact that a restricted vocabulary set tends to be used by the online news sources they selected.

The purpose of the second experiment was to determine how well the system classified unseen news articles. Using 549 financial news articles out of 1,168 downloaded articles, the "domain experts with vote entropy" system correctly classified 433 of those articles as 'good', 'neutral', or 'bad' i.e. the classifier achieved 79% accuracy on average. Only one of the 549 articles had been manually labelled as 'uncertain'; all other articles had been labelled 'good', 'neutral', or 'bad'. The same data was used with the naïve Bayes; the result there was 65%. Strengths of this study include their use of multiple evaluation metrics (random guessing, comparison with the results from the most frequent class, and classification accuracy). They also employed various feature selection methods.

van Bunningen (2004) used pharmaceutical news articles downloaded from the Dow Jones Newswire to predict market reactions to those news articles. He interviewed traders, who told him that news relating to approval and disapproval of patents tends to be highly price sensitive. He chose pharmaceutical companies because many of their news articles relate to approval or rejection of patents. Whilst he also said that trader reactions to pharmaceutical companies tend to be more predictable than reactions to IT companies, this claim was not substantiated in his study.

In order to isolate the effect of news articles from other market effects, van Bunningen decided to use the market average return; this compares the stock price of a company with the stock prices of other companies at the same moment. Other methods that could have been used include the mean adjusted returns, whereby "returns are compared with the returns from the same company on other days" and the market and risk adjusted returns, which also consider "other high risk assets" (*ibid*, p.18). van Bunningen broadly defined a trend as "the difference between two closing prices on two consecutive days" (*ibid*, p. 64). Articles that were released between those two days were then aligned with that trend. Other studies defined trends and patterns in terms of *high frequency intraday data* (e.g. Ederington and Lee 1993; Goodhart and O'Hara 1997).

He underpinned his research with three hypotheses: (1) news article analysis has significant advantage above trend analysis, (2) news articles have an immediate influence on the stock market, and (3) the predictable influence of news articles is only directly after it arrives and therefore one must react as quickly as possible and ensure that news articles are correctly aligned with reactions. In relation to hypothesis (3), the author only used closing prices but admitted that many would argue that one should use intra-day prices rather than closing prices (see Section 2.4 for a discussion on market reactions to events). He also assumed that all returns are abnormal and that all companies are equally affected by external factors; in reality, this would not be the case. He did not cater for transaction, spread, or inflation costs.

He chose as a 'surge' the 50% most steepest positive trends. The 50% most steepest negative trends were used as a 'plunge'. His dataset comprised twelve months of

closing prices and approximately 14,000 documents for 23 companies. The entire data set was split into seven training sets coupled with seven test sets; the smallest training set spanned approximately three and a half months and the accompanying test set spanned approximately two months. The largest training set spanned almost twelve months and the accompanying test set spanned approximately one month.

He used template filling (e.g. see MUC 1997) to extract features from the news articles and IDF to attach weights to the features. Only features that appeared a minimum of three times were used. Example features included *it generate(10) sale(0)*, *#\$ThisCompany report(10) loss(10)*, and *#\$stock trade(10) #\$price(0)*. Some patent-specific features included *#\$Company get #\$PatentApproval* (sic) and *#\$Company settle charge*. He then used a support vector machine (SVM) to classify each document as a 'surge', a 'plunge', or a 'not relevant' document. Unlike Lavrenko et al (2000) (see Section 3.2), van Bunningen decided not to use categories for 'slight plunges' and 'slight surges', as he did not think these categories were necessary.

He performed a number of tests with the data, including tests with all available articles, tests with a selection of articles that he considered to be more influential, and tests using headlines only. When he used all available data, the system classified 40.67% of the documents as 'surges' whereas only 32.07% were in fact 'surges'; it also classified 37.28% of the documents as 'plunges' but only 33.93% were in fact 'plunges'. Likewise, when he tested with only a selection of articles, the system classed 38.51% as 'surges' and 41.35% as 'plunges' whereas only 32.16% and 33.96% were in fact 'surges' and 'plunges'. When he only used headlines, 55.76% and 48.84% were classified as 'surges' and 'plunges' respectively, but only 32.07% and 33.93% were in fact 'surges' and 'plunges'.

He also found that the majority of articles tended to be classified as 'not relevant'. For example, in the first data set that comprised all available data, 428 articles were classified as 'not relevant' when they were in fact 'plunges', whilst only 36 were correctly classified as 'plunges'. Likewise, 493 articles were predicted as 'not relevant' when they were in fact 'surges', whilst only 64 were correctly classified as 'surges'. He proposed that it was possible that those articles may have contained

features that occurred less than three times in the document collection and therefore those features had never been seen by the system before. He also suggested that there might not be a high correlation between the news articles and the stock prices and that these factors, coupled with the fact that the 'plunges' in the S&P 500 index were the largest, might have contributed to the high number of 'not relevant' classifications. To identify useful features and improve the classification, he recommended using information gain and more domain knowledge, with intraday data. Some strengths of this study include their use of multiple evaluation metrics (random chance, classification accuracy, and precision and recall). The researchers did not, however, compare the results with those from baseline approaches (e.g. bag-of-words) or a trading strategy.

### **3.5 Financial Ratios and Variables**

Koh and Low (2004) looked at financial ratios for 330 companies, derived from various electronic sources (the COMPUSTAT tapes, Moody's Industrial and OTC Manuals, and the SEC's 10-K Reports), to predict going concern companies. Using a dataset of 165 going concern and 165 non-going concern companies, they evaluated the usefulness of decision trees, neural networks, and logistic regression to predict going concern and non-going concern status. They found that the decision tree models outperformed the other two models when the same sample was used for training and testing; decision trees achieved a classification accuracy rate of 97.39%, whereas the accuracy for the neural network and logistic regression models were 95.65% and 95.22% respectively.

They also examined the sensitivity of each type of model to changes in the independent variables and found that the total liabilities to total assets (TLTA) and retained earnings to total assets (RETA) variables were the most important variables, when predicting going concern status, for each of the three models. They used the Berry and Linoff (1997) method, which involves analysing prediction changes based on the average, minimum and maximum values of each independent variable. They found that the impact of the TLTA and RETA variables was found to be in the expected direction. When they used a validation sample, decision trees still

outperformed the other two models, by achieving a classification accuracy of 95%. The accuracy for logistic regression and neural networks was 94% and 91% respectively. These high results seem to imply that financial ratios are very good predictors of going concern companies, but unfortunately the authors did not report on the statistical significance of these results. They identified some limitations of their own study, which included the fact that it might be more interesting and useful if the models were able to select the most predictive variables. Also, they suggested that one should consider the costs of any misclassifications and the relative proportion of going concern and non-going concern companies used in the experiments.

Qi (1999) used nine microeconomic and macroeconomic variables as well as a neural network to predict stock returns on the S&P 500. The microeconomic variables were 'dividend yield' and 'earnings-price ratio'. The macroeconomic variables included the 'one-month Treasury-bill rate' and the 'year-on-year rate of inflation'. Qi compared the performance of the neural network (NN) to linear regression (LR) and found that the performance of both models depended not only on the out-of-sample test period, but also on the performance measures used (e.g. root-mean-squared error or the Pearson correlation coefficient).

When he tested the statistical significance of the results he found that the neural network had smaller squared forecast errors than linear regression but that these were not statistically significant. Qi also determined that, despite the level of forecasting errors in both models at different time periods, both the neural network and linear regression have some market timing ability, which means they can still be used by investors to generate wealth. He used the Pesaran Timmermann (PT) nonparametric test to determine this (see Pesaran and Timmerman 1992 for a discussion of the PT test). In particular, they found that using a switching portfolio and the NN model yielded higher profits with lower risks than using a buy-and-hold market portfolio and LR. However, another author, Racine (2001), who attempted to replicate Qi's results, cautioned against relying on that trading strategy because he found the opposite was true i.e. that using a switching portfolio and the NN model yielded lower profits with higher risks than using a buy-and-hold market portfolio and LR.

Qi tested the relative predictive power of both models using an encompassing test and found that the nonlinear neural network had significant explanatory power. The test also showed that the information contained in the LR forecasts was simply a subset of the information contained in the NN forecasts. To ensure robustness, he tested the profitability of various transaction costs and different types of portfolios over different decades; the portfolios were market, bond, and switching portfolios. Using zero transaction costs, for example, the neural network switching portfolio yielded higher profits with lower risks for each of the decades, than using a 'buy-and-hold' market portfolio or the linear regression switching portfolio; the linear regression switching portfolio also yielded higher profits than the 'buy-and-hold' market portfolio.

Kryzanowski et al (1993) attempted to predict future returns by looking at industry and company financial ratios, as well as seven macroeconomic variables, obtained from financial statements and business publications. The five industry ratios ('gross profit margin', 'current ratio', 'net profit margin', 'return on equity', and 'total debt-to-equity ratio') were used to set a benchmark for each company. The 14 financial ratios included the 'gross profit', 'total asset turnover', and 'debt ratio' for each company. These were broadly categorised as profitability, debt, or liquidity and activity ratios. The seven macroeconomic variables included 'industrial production', 'gross domestic product', and the 'consumer price index'. In total, 88 variables were used; the variables also included the volatility of each of the ratios for each company and the volatility of each ratio for the industry in general.

They conducted two experiments using the Boltzmann Machine (BM) algorithm for pattern classification. A BM "is a kind of ANN that uses simulated annealing to set the states of the neurons during both the weight-learning and function-computing stages of its operation" (*ibid*, p.22). Simulated annealing is "a discrete optimization method that uses a gradually decreasing amount of random noise while searching some problem space for a globally optimal solution" (*ibid*, p.22).

In the first experiment, 40 training cases and 42 verifying cases were used to classify 149 test cases as having a 'positive' or 'negative' return. In this two-way

classification, the BM correctly classified 66.4% of the 149 test cases or 71.2% when the 11 cases which it could not decide on, were omitted (leaving just 138 cases). Whilst the negative cases had many more errors than the positive cases, it is important to note that there were significantly more negative cases (125) than positive ones (24). They also found that one year (1989) yielded more errors than the two previous years. When they analysed the errors more closely, they discovered some unusual findings. For example: one case, which actually resulted in a 150% return, was predicted by the system to yield a negative return. Another case, which actually resulted in a -73% return, was predicted by the system to yield a positive return.

In the second experiment, they used 39 training cases and 42 verifying cases to classify 149 cases as positive, neutral, or negative. In this three-way classification, the BM correctly classified 45.6% of the 149 cases. However, the BM incorrectly classified 46 out of 47 neutral cases as either 'positive' or 'negative'. They did suggest, however, that despite these promising results, there was a need for experiments with a larger number of cases and different time frames, to ensure robustness.

To evaluate the system, they compared the results from both experiments with random guessing. In the first experiment (the two-way classification), the accuracy was 66.4% as opposed to 50%. In the second experiment (the three-way classification), the accuracy was 45.6% as opposed to 33.3%. Unfortunately, they did not report on the potential profitability of the system.

Lam (2004) predicted the rate of return on common shareholders' equity for 364 S&P companies for the ten-year period 1985 -1995 by integrating fundamental and technical analysis (see Section 2.1 for a discussion on the two types of financial analyses). She used 120 'high-return' companies, 122 'medium-return' companies, and 122 'low-return' companies in the dataset.

To evaluate the performance, she compared the average return from all the companies classified by the system as 'high return' companies, with maximum and minimum benchmarks, rather than use the percentage of correctly classified cases as the

performance indicator. The maximum benchmark was the average return from the top one-third returns in the market and the minimum benchmark was the market average in the test set.

Using recommendations from previous studies including Kryzanowski et al (1993), Lam used 11 macroeconomic variables extracted from the Compustat file and 16 microeconomic financial statement variables (referred to as 'microeconomic variables' from now on) extracted from the industrial annual file. Macroeconomic variables included 'government spending/gross domestic product', the 'consumer price index', and the 'purchase price of crude oil'. Microeconomic variables included 'net income/net sales', 'dividend per share', and 'common shares traded'. Using a neural network, she conducted five experiments (see Table 3.11).

	<i>Number of Years Financial Data</i>	<i>Number of Variables Used</i>	<i>Number of Sets of Training and Test Samples</i>	<i>Number of Companies in each Training and Test Set</i>
<i>Experiment one</i>	1	16 micro variables	9	Training set: 364 Test set: 364
<i>Experiment two</i>	2	32 micro variables	8	Training set: 364 Test set: 364
<i>Experiment three</i>	3	48 micro variables	7	Training set: 364 Test set: 364
<i>Experiment four</i>	3	16 micro variables 11 macro variables	7	Training set: 1092 Test set: 364
<i>Experiment five</i>	1	16 micro variables	9	Training set: 364 Test set: 364

Table 3.11: Results of neural network experiments using different years, variables, training, and test sets (Lam 2004).

In experiments one to three, the neural network was unable to meet the maximum benchmark; only the minimum benchmark was "consistently and significantly" outperformed (*ibid*, p.567). In experiment four, the NN was able to outperform the minimum benchmark, but only marginally. The maximum benchmark was not met. In experiment five, Lam used the same data and number of training and test sets as in experiment one. However, she then used the GLARE algorithm to extract classification rules from the neural network (see Appendix A of Lam 2004 for an overview of the input/processing/output of GLARE). One rule (rule 5) achieved an

average return that was 53% better than the average return from the original neural network; this was also the best average return among all experiments and almost met the maximum benchmark. As a result, Lam recommends the use of post-processing when pre-processing is impractical. Apart from improving performance, Lam maintains that post-processed rules are useful for explaining prediction logic to investors.

### **3.6 Summary**

In this chapter, we presented studies that analysed single words, keyword records and phrases, and financial ratios and variables in financial documents. Whilst eighteen studies involved the analysis of single words, four of these studies also experimented with keyword records and phrases. Seven studies used only keyword records and phrases. Two studies used single words with financial ratios and variables and six used financial ratios and variables only.

In terms of data sources, twenty-three studies used online news stories or messages but only five studies used on- or off-line statements or reports. Prediction goals varied—the majority of studies (twenty-three) involved the prediction of prices, directions, or trends. Three studies involved the prediction of market or message sentiment and two studies involved other prediction goals (i.e. going concern companies and the volatility of companies or markets).

A number of different methods were used. Five studies used rule induction methods, seven used neural networks, and three used both rule induction methods and neural networks. Five studies used support vector machines and nine used Bayesian methods. One study used rule induction methods and Bayesian methods and two used support vector machines and Bayesian methods. Twenty-three studies used 'other methods', which included statistical methods, language modelling, multiple discriminant analysis, k-nearest neighbour, and genetic algorithms.

With regards single word features, support vector machines and Bayesian methods were used more frequently (used five and eight times respectively) than rule induction

methods and neural networks (both used once only). 'Other methods' were also used but often to compare with the aforementioned methods. With regards keyword records and phrases, rule induction methods were used five times whereas neural networks, support vector machines, and Bayesian methods were used three, three, and two times respectively. With regards financial ratios and variables, neural networks were used four times, rule induction methods and Bayesian methods were both used once, and support vector machines were not used.

Table 3.12 recaps the methods adopted in the various studies (also in Table 3.1) but also highlights the range of evaluation metrics used and the research considerations, which varied greatly between studies. Feature selection methods such as TF\*IDF are categorised as 'other research considerations' except where they were used as a baseline approach, in which case they appear as an evaluation metric under 'Compared method(s) with baseline approach(es)' (see, for example, Lavrenko et al., 2000).

Author(s)**	Page Number	Methods					Evaluation Metrics					Research Considerations		
		Decision trees and rule induction methods	Neural network methods	Support vector machines	Bayesian methods	Other methods	Random guessing/ chance	Compared method(s) with baseline approach(es)*	Classification/ prediction accuracy or hit rate analysis	Precision, recall, and/or F-measure	Adopted a trading strategy (e.g. buy-and-hold)	Cross-validation	Covered for transaction costs, inflation, and/or spread	Other research design considerations**
Das and Chen (2001)	50				*	*	*		*					
Swales and Yoon (1992)	53		*			*			*					
Schumaker and Chen (2006)	54			*		*		*	*		*	*		*
Mittermayer and Knolmayer (2006a)	56			*		*		*	*		*	*		*
Fung et al (2005)	60			*		*		*	*		*			*
Antweiler and Frank (2004)	61			*	*				*			*		*
Antweiler and Frank (2006)	62				*				*					*
Gidófalvi (2001)	63				*	*	*		*	*				*
Liu et al (2006)	65					*					*	*		*
Lavrenko et al (2000)	67				*	*	*	*	*	*	*	*		*
Koppel and Shtrimerberg (2004)	69	*		*	*				*	*		*		
Kroha and Baeza-Yates (2004)	71				*	*			*					
Kroha et al (2006)	73				*	*			*					
Tetlock (2007)	74					*	*				*	*		*
Tetlock et al (2008)	75					*					*	*		*
Engelberg (2008)	76					*					*			*
Henry (2008)	79					*			*					*
Loughran and McDonald (2011b)	80					*		*	*		*	*		*
Permunetilleke and Wong (2002)	85	*	*			*	*		*			*		*
Wüthrich et al (1998)	87	*	*			*			*		*			
Thomas and Sycara (2000)	89					*					*			*
Cho et al (1999)	90					*	*		*		*	*		*
Thomas (2003)	92	*							*	*	*	*	*	*
Seo et al (2004)	94				*	*	*	*	*					*
van Bunnigen (2004)	96			*		*	*		*	*				*
Koh and Low (2004)	98	*	*			*			*					*
Qi (1999)	99		*			*			*		*	*		*
Kryzanowski et al (1993)	100		*				*		*					*
Lam (2004)	101		*								*			*
<b>Slattery (2012)</b>		*		*	*		*	*	*			*		*

Table 3.12 Methods, evaluation metrics, and research considerations used for the automatic analysis and classification of financial documents.

\* Examples of baseline approaches include the most frequently-occurring class or naive bayes bag-of-words.

\*\*Other research design considerations include feature selection, robustness analysis, statistical significance, or various event windows.

Because the features, datasets, and methods varied so widely, it is not possible to identify a single best result. However, we have identified what we believe to be three of the best studies, based on our evaluation of the rigor of their research. The strengths of each of these studies are briefly summarised in Table 3.13. Some limitations are also summarised, in the interest of completeness.

<i>Author(s)</i>	<i>Strengths</i>	<i>Limitations</i>
Schumaker and Chen (2006)	<ul style="list-style-type: none"> <li>• Examined three document representations (bag-of-words, noun phrases, and named entities).</li> <li>• Used SVM for all three representations and compared three results (closeness, directional accuracy, and profitability) with the equivalent results from linear regression.</li> <li>• Considered the cash outlay/ investment required to make various returns.</li> <li>• Performed 10-fold cross validation and used statistical techniques to measure robustness.</li> </ul>	<ul style="list-style-type: none"> <li>• Relatively small dataset.</li> <li>• Only used S&amp;P companies (could also be seen as a strength).</li> <li>• Did not investigate other machine learning approaches.</li> <li>• Did not consider transaction, inflation, or spread costs.</li> </ul>
Mittermayer and Knolmayer (2006a)	<ul style="list-style-type: none"> <li>• Used a thesaurus of single words with a bag-of-words model.</li> <li>• Employed various feature selection techniques (inverse document frequency (IDF), collection term frequency (CTF), and information gain (IG)).</li> <li>• Employed various algorithms (linear and non-linear SVMs, k-nearest neighbour (k-NN), and Rocchio).</li> <li>• Performed 10-fold cross validation.</li> <li>• Developed a trading engine and used a two-sided early exit strategy.</li> <li>• Compared results with random guessing.</li> <li>• Examined prediction accuracy.</li> <li>• Examined harmonic mean of macro average precision and recall.</li> <li>• Performed a robustness analysis by adjusting feature selection parameters, size</li> </ul>	<ul style="list-style-type: none"> <li>• Excessive filtering of documents.</li> <li>• Did not consider transaction, inflation, or spread costs.</li> </ul>

	of feature set, document representation, and classifier.	
Lavrenko et al (2000)	<ul style="list-style-type: none"> <li>• Used Bayesian language modelling and linear regression.</li> <li>• Used external relevance assignments/ judgements.</li> <li>• Evaluated performance using detection error tradeoff (DET) curves.</li> <li>• Compared the performance of their method (LM) with the vector-space method.</li> <li>• Implemented a basic trading strategy/ simulation.</li> </ul>	<ul style="list-style-type: none"> <li>• Ignored temporal ordering.</li> <li>• Only considered volatile stocks.</li> <li>• Did not consider transaction, inflation, or spread costs.</li> <li>• Unlimited and unrealistic funds during the simulation.</li> </ul>

Table 3.1 3: Three rigorous studies determined by the range of evaluation metrics used and the robustness of their research methods.

As we will discuss in Chapter 6, we examined the performance of two machine learning algorithms. We also used a combination of evaluation metrics, including directional accuracy on unseen data, random guessing/chance, and comparison with baseline approaches. Whilst we did not consider transaction, inflation, or spread costs or implement a trading strategy, we do recommend future work in this area in Chapter 7. But before we examine our methodology and results, we need to present the background and rationale to our study. In Chapter 4, we will examine the data sources we chose (Form 8-K disclosures) and why we chose them. We will also examine the wider price dynamics in which our 8-Ks were filed, by looking at various event windows. We will then describe how we identified financial event phrases within 8-Ks.

# **Chapter 4: Events in Form 8-K Disclosures**

## **4.1 Outline**

In the previous chapter, we examined studies that used single words, keyword records and phrases, and/ or financial ratios and variables for the automatic content analysis of financial documents. In this chapter, we begin to present the rationale for our study. In Section 4.2, we briefly look at the SEC legal requirements for the filing of corporate disclosures. We also examine Form 8-K disclosures in detail (hereafter referred to as 8-Ks), as we use 8-Ks in this study. Finally, we look at some of the features of the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system, which hosts 8-Ks. In Section 4.3 we examine the datasets used in this study and in Section 4.4, we examine the causal effect of 8-Ks on prices, by considering the wider price dynamics in which our 8-Ks were filed. In Section 4.5 we discuss how we identified the financial event phrases (FEPs) that were used to recognise events in 8-Ks. Finally, in Section 4.6, we provide a summary of the chapter.

## **4.2 Background to Corporate Reporting on the EDGAR System**

The Securities and Exchange Commission (SEC), headquartered in Washington DC, enforces a number of laws to ensure that investors and markets have access to all the information that is needed to trade in a fair, orderly, and efficient manner (SEC 2010b). Two particularly important laws relating to the filing of corporate information are the Securities Act of 1933 and the Securities Exchange Act of 1934.

The Securities Act of 1933 requires companies to ensure that investors have access to all relevant information concerning the public sale of securities. The Act covers issues such as deceit, misrepresentations, and other fraud relating to the sale of securities (SEC 2010b). This Act also requires the majority of U.S. companies to register their securities with the SEC. The Securities Exchange Act of 1934, which led to the establishment of the SEC, requires companies to ensure that investors have access to all relevant information about publicly-traded and over-the-counter (OTC)

securities. Also, as part of this Act, companies with assets worth more than \$10 million and whose securities are held by more than 500 owners, are required to file periodic forms including 10-Qs (quarterly reports), 10-Ks (similar to the annual reports), and 8-Ks (current or material event reports). There are over 100 different form types and each relates to specific events which must be disclosed to investors and other relevant parties. Each form type has its own legal requirements also; for example, the filing deadlines vary for each form type.

Our thesis is concerned with 8-Ks and we chose this form type for two reasons. Firstly, an initial evaluation of form types revealed that companies are required to file 8-Ks whenever material events or corporate changes which are of importance to investors and other relevant parties arise. Secondly, we had access to closing prices within a three-day window of corporate filings, so it made sense to use 8-Ks as they must be filed within a reasonably short time frame after a triggering event. An 8-K contains more current information than is found in a 10-Q or 10-K, and is therefore called a 'current report'. When amendments are made to 8-Ks they are called 8-K/As but we did not examine these.

Initially, there were nine reportable events in 8-Ks (see Table 4.1). Some events were clearly of relevance to investors (e.g. item 3: bankruptcy) but others were of less significance (e.g. item 8: change in fiscal year). Also, as we will discuss shortly, companies were only required to file some items *after* the event took place, which probably diluted the impact of those 8-Ks on share prices, as the information had already reached the market by the time the 8-K was filed. However, in June 2002, the SEC put forward a proposal (SEC 2002) to increase the number of reportable events to 22. The SEC also proposed that the filing deadline be shortened from five business or fifteen calendar days (depending on the event) to two business days (or four days for companies who filed a Form 12b-25), to increase the timeliness—and therefore the usefulness—of 8-Ks to investors. Due to the large number of comment and complaint letters they received, they did not implement this proposal.

Item 1	Changes in Control of Registrant
Item 2	Acquisition or Disposition of Assets
Item 3	Bankruptcy or Receivership
Item 4	Changes in Registrant's Certifying Accountant
Item 5	Other materially Important Events
Item 6	Resignation of Registrant's Directors
Item 7	Financial Statements and/ or Exhibits
Item 8	Change in Fiscal Year
Item 9	Sales of Unregistered Equity Securities/ Regulation FD Disclosure

Table 4.1: Initial listing of reportable event items in Form 8-Ks (SEC 2002).

In July 2002, Congress passed the Sarbanes-Oxley Act (SEC 2002)<sup>28</sup>. Amongst other things, this Act required public companies to disclose information that was of material interest to investors and other relevant parties, on a rapid and current basis. Following on from this Act, and having taken on board many of the comments received after the 2002 proposal, the SEC officially amended the 8-K requirements in March 2004 (SEC 2004a). Eight new items were added to the list of events that must be reported, two items were transferred from the periodic reports (10-Qs and 10-Ks), and two existing items were modified. Items were reorganised and renumbered and similar items were grouped (see Table 4.2). Also, the time period within which such filings must be made was shortened to four days after the triggering event. The changes took effect in August 2004.

Section 1 Registrant's Business and Operations	Item 1.01	Entry into a Material Definitive Agreement
	Item 1.02	Termination of a Material Definitive Agreement
	Item 1.03	Bankruptcy or Receivership
Section 2 Financial Information	Item 2.01	Completion of Acquisition or Disposition of Assets
	Item 2.02	Results of Operations and Financial Condition
	Item 2.03	Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant
	Item 2.04	Triggering Events that Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement

<sup>28</sup> <http://www.sec.gov/about/laws/soa2002.pdf>

	Item 2.05	Costs Associated with Exit or Disposal Activities
	Item 2.06	Material Impairments
Section 3 Securities and Trading Markets	Item 3.01	Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing
	Item 3.02	Unregistered Sales of Equity Securities
	Item 3.03	Material Modifications to Rights of Security Holders
Section 4 Matters Related to Accountants and Financial Statements	Item 4.01	Changes in Registrant's Certifying Accountant
	Item 4.02	Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review
Section 5 Corporate Governance and Management	Item 5.01	Changes in Control of Registrant  Departure of Directors or Principal Officers; Election of Directors; Appointment of Principal Officers
	Item 5.02	
	Item 5.03	Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year
	Item 5.04	Temporary Suspension of Trading Under Registrant's Employee Benefit Plans
	Item 5.05	Amendments to the Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics [See Table 4.3 for later amendments to this Section]
Section 6	[Reserved]	[See Table 4.3 for later amendments to this Section]
Section 7 Regulation FD	Item 7.01	Regulation FD Disclosure
Section 8 Other Events	Item 8.01	Other Events
Section 9 Financial Statements and Exhibits	Item 9.01	Financial Statements and Exhibits

Table 4.2: Amended listing of reportable items in Form 8-Ks (SEC 2004a).

The increase in the number of material events that needed to be filed caused concern to many companies, some of whom had rarely filed 8-Ks because they had considered the majority of their events to be of little interest to investors. Prior to 2004, companies only had to report changes of control or the acquisition or disposition of assets *after* the transaction was completed. As a result, the 8-K information was probably not very useful, as the information would have already reached the market through other sources. However, the new SEC requirements meant that a company

had to file an 8-K *as soon as* it entered into an agreement and again after closing the agreement. The filing deadline changes, in particular, brought with them an increasing concern that events that might not seem material, might indeed turn out to be so. However, for seven of the event types, failure to make a filing within four days of the event would not make the company liable *provided* the company reported the event(s) in a 10-Q for the quarter in which the event(s) occurred.

According to Bernstein (2004), the increase in the number of events that needed to be reported brought with it a reduction in the value of 8-Ks, as companies scrambled to file them by the deadline; also, the stricter rules meant that many 8-Ks contained immaterial as well as material events. Bernstein argued that the majority of events that must be reported in 8-Ks would appear in press releases almost immediately afterwards anyway; therefore, there was no real advantage to imposing such a short filing deadline on companies. However, a more recent study by Lerman and Livnat (2009), which examined how informative the new 8-Ks are, found that the new guidelines provided investors with more timely access to relevant information (see Section 2.4 for a discussion). See Table 4.3 for event items that were added between 2005 and 2011.

Section 5 (effective November 2005)	Item 5.06	Change in Shell Company Status.
(effective February 2010)	Item 5.07	Submission of Matters to a Vote of Security Holders.
(effective January 2011)	Item 5.08	Shareholder Director Nominations.
Section 6 Asset-Backed Securities (effective March 2005)	Item 6.01	ABS Informational and Computational Material.
	Item 6.02	Change of Servicer or Trustee.
	Item 6.03	Change in Credit Enhancement or Other External Support.
	Item 6.04	Failure to Make a Required Distribution.
	Item 6.05	Securities Act Updating Disclosure.
(effective January 2011)	Item 6.10	Alternative Filings of Asset-Backed Issuers.

Table 4.3: Subsequent amended listing of reportable event items in Form 8-Ks (SEC 2004b; SEC 2005; SEC 2009a; SEC 2011).

Apart from the relatively recent study by Lerman and Livnat, few studies have examined the impact of 8-Ks on returns. As we discussed in Chapters 2 and 3, most corporate news prediction studies focused on *specific event types* or they used online news stories or messages, rather than 8-Ks. Also, as none of the other studies used *keywords and events* in 8-Ks as prediction features, this makes our approach unique.

This thesis uses three datasets of 8-Ks. The first dataset we obtained related to specific types of companies that filed over a five-year period in the mid to late 1990s and were therefore only required to file a limited number of event types (see Items 1-9 in Table 4.1). We used this initial dataset for some preliminary experiments and to develop the grammar of financial event phrases (see Section 4.4). The second and third datasets related to four-year periods but to a much broader spectrum of companies. Also, the third dataset relates to the post 2004 SEC changes, so these disclosures should contain many more types of events. See Section 4.3 for a thorough description of the three datasets used.

In terms of the structure of 8-Ks, they consist of submission header information and concatenated documents. The submission header typically contains the following information (Keane 2010):

- Accession number<sup>29</sup>.
- Form type.
- Filed as of date (filing date).
- Company conformed name.
- Central Index Key (CIK)<sup>30</sup>.
- Standard Industry Classification (SIC)<sup>31</sup>.
- IRS number.
- Fiscal year end.
- Business address.

---

<sup>29</sup> The accession number is a unique number generated by EDGAR, comprising the CIK, the year, and a six-digit sequence number. The accession number also appears in the URL for the filing.

<sup>30</sup> The CIK is a unique identifier assigned by the SEC to each person or entity who files disclosures with the SEC. In 1997, the North American Industry Classification (NAICS) system officially replaced the SIC, although the SIC is still used in EDGAR filings.

<sup>31</sup> The SIC is a code used to classify each filer by industry. For example, the 6021 SIC code refers to companies in the Financial Services industry.

The main body of the 8-K comprises applicable event items, other concatenated documents, and signatures. Companies are permitted to incorporate event information by reference to sources such as press releases or other statements, provided those sources are published within the prescribed period for filing a 8-K and they include copies of those other sources as exhibits in the 8-K (SEC 2010a). Companies are obligated to ensure the reports are informative and not misleading. In 1998, the SEC published 'A Plain English Handbook' (SEC 1998) to help corporate lawyers and authors write clearer and more informative disclosures.

The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, which was officially launched in 1994, is an online archive of all forms filed electronically with the SEC. Prior to 1994, disclosure filings had to be made by hard copy. By May 1996, all public domestic companies were required to make their filings on EDGAR, except in the case of a hardship exemption. EDGAR now accepts and disseminates 4,000 to 12,000 live public submissions per day during peak periods; put another way, EDGAR receives up to 2GB of data in a single business day (Keane 2010). The size of disseminated submissions can vary from 1KB to 100MB but the average submission size is around 115KB (Keane 2010).

As we have already pointed out, our initial dataset relates to the mid to late 1990s. At that time, filers were restricted from including active content, external references, nested table tags, and any other tags that were not supported by the EDGAR system. The majority of the filings in that dataset comprised of limited tags, apart from the required header tags and closing `</TEXT>` and `</DOCUMENT>` tags which were automatically inserted by the EDGARLink software, which was used to prepare and upload the 8-Ks. However, since the late 1990s, a number of major improvements have been made to EDGAR. Nowadays, concatenated documents within submissions can be formatted in ASCII/SGML, HTML, GIF, JPG, PDF, XML, or XBRL and they can be uploaded via the Internet or Secure Shell (SSH) (Keane 2010). In 2011 and 2012, the three volumes of the EDGAR filer manual were updated for EDGAR users to provide detailed information on the application and filing processes<sup>32</sup>. The

---

<sup>32</sup> <http://www.sec.gov/info/edgar/edmanuals.htm>

EDGARLink software and submission templates provide filers with guidelines on how to format their disclosures. Recent developments in XBRL reporting have resulted in a significant improvement for EDGAR users and filers. XBRL tagging, also known as interactive disclosure technology, facilitates users to more easily compare information, whilst also minimising the burden on filers. The SEC believes that users will no longer be forced to pay third-parties to repackage the financial data before they can analyse it. One major advantage of XBRL data is that it can be downloaded directly to spreadsheets and analysed in a variety of off-the-shelf products (SEC 2009). XBRL requirements were phased in from 2008 but were officially adopted in a Final Rule in 2009 (SEC 2009b). Filers are now required to file the interactive XBRL data as an exhibit to their traditional filings and the traditional filings will continue to be offered to users. These recent changes have been designed to go hand-in-hand with plans to eventually replace the EDGAR system with the Interactive Data Electronic Applications (IDEA) system.

It currently takes no longer than two minutes from the receipt of an electronic submission by EDGAR to the dissemination of a validated and reassembled version via the EDGAR Public Dissemination Service (PDS) and SEC website (Keane 2010). The SEC still accepts paper submissions but these are entered into EDGAR by the SEC. Abbreviated versions of the paper submission are then disseminated to the public via another service.

A variety of search options are now available on EDGAR. General purpose searches include searching for companies and filings by company name, ticker symbol, Central Index Key (CIK), or Standard Industry Classification (SIC); searching the most recent filings—both paper and electronic; and searching for current events within the past five days. It is also possible to perform a full-text search of filings (including attachments) for the past four years. Special-purpose searches include searching for mutual funds filings and retrieving prospectuses for specified mutual funds. EDGAR users can also use File Transfer Protocol (FTP) to access indices of archived data sorted by company, form type, and CIK (known as the master index).

## 4.3 Datasets

As mentioned previously, we used three datasets. The first was a dataset of 8-Ks in the 7372 SIC, spanning from 1994 to 1998. The 7372 SIC refers to ‘Services-Prepackaged Software’ and includes companies that develop and sell applications software, operating systems, and other software utilities. We initially obtained data for this industry purely because it was related to the field of computing. However, to ensure transferability to other industry domains and to make certain that we held no bias towards small or large firms, we later randomly chose 50 companies from the Standard & Poors (S&P) 500 index for the second and third datasets. We will now discuss the features of each dataset.

### 4.3.1 The 7372 Dataset

We initially obtained the filing dates and URLs for 556 8-Ks filed between January 1994 and September 1998. These 8-Ks all related to companies in the 7372 SIC domain. For each of the 8-Ks, we also received the closing stock price data for days  $t-1$ ,  $t$ , and  $t+1$ , where  $t$  was the day of the EDGAR filing. The percentage share price return around the announcement date (from here on referred to as the annret) was then calculated as follows:  $((P_{t+1})/(P_t))-1$ , where  $P$  stands for the closing share price. The annrets ranged from -1.00 to +0.88 for these 556 8-Ks.

Like van Bunningen (2004)<sup>33</sup>, we used closing prices rather than intraday prices as only closing prices were readily available. Also, it was not always possible to determine the specific time of the day when filings were submitted to the SEC and subsequently posted online, so it was best to use a consistent pricing method. Whilst Mittermayer and Knolmayer (2006b) argue that closing prices should not be used as the market reacts very quickly, and Andersen and Bollerslev (1998b) argue that closing prices that are similar to the previous opening prices may disguise intraday volatility, MacKinlay (1997) was not convinced of the net benefit of high-frequency intraday data. We worked on the assumption that closing prices within a day of an

---

<sup>33</sup> van Bunningen (2004) compared the closing price of a company with the closing price of the market in general, to isolate the effects of the news article on the price. He referred to this as the market adjusted return, as discussed in Brown and Warner (1985). We, on the other hand, used a more simplified return model which compared the closing prices of a company around the filing release date.

event would still capture the effect of the event but we will discuss the consequences of this approach in Chapter 7.

We chose a three-day window (days  $t \pm 1$ ) for a number of reasons. Firstly, we assumed that reactions to 8-Ks—as opposed to news headlines—do not always occur immediately and may take up to a day (see Section 2.2 for a discussion of the Efficient Markets Hypothesis). In other words, we assumed that the market is not fully efficient (see Section 1.3 for our list of assumptions). Secondly, as SEC filings are sometimes posted after trade has ceased on day  $t$ , the effect of an 8-K might not be seen until day  $t+1$ . Also, in his 2003 study, Griffin (2003) reported there can be a short lag between when an EDGAR filing is made and when a filing appears on the public site. However, because the filing deadline was reasonably generous prior to 2004 (five business or fifteen calendar days), it is quite possible that the event information had already been leaked, by the time our 8-Ks were filed. On that note, Carter and Soo (1999) found evidence of a *limited* market response to 8-Ks and that timely filing was the single most important factor influencing the market relevance of the 8-K. As discussed in Section 2.4, Lerman and Livnat examined the market relevance of 8-Ks filed *since* the 2004 rule changes and found that the newer 8-Ks provide investors with more timely access to relevant information. We addressed the issue of the market relevance of 8-Ks to a certain extent, by obtaining data that was filed after the 2004 rule changes (the third dataset, which will be discussed shortly). In Section 4.4, we will also discuss the causal effect of 8-Ks in various windows.

Antweiler and Frank (2006) suggested using windows greater than three days for robustness as they found that the returns varied depending on the length of window used (see Section 3.2 for a discussion of their study). Tetlock (2007) also experimented with different window sizes and found that some negative effects were reversed within a week (see Section 2.4). Asthana and Balsam (2001) used a five-day window (days  $t-1$  to  $t+3$ ) but reported similar results with three-day (days  $t-1$  to  $t+1$ ) and seven-day (days  $t-1$  to  $t+5$ ) windows. Whisenant et al (2003) used three- and seven-day windows but suggested that if leakage of information occurs before an 8-K filing is made, then the use of a three-day window could potentially “underestimate the information content of the disclosures” (p.185). Griffin (2003)

examined 10-K and 10-Q filings and found a significant market response to 10-Ks within three days after the filing date, particularly for smaller firms, firms with less institutional ownership, on days when there is a high level of filings, and when the filing arrives after the due date. The response to 10-Q filings was broadly similar, although the response was more elevated after 10-K filings; they also found that the window of the significant response was within two days for 10-Qs, compared to three days for 10-Ks. Whilst the Griffin study did not examine 8-Ks, we believe the market responses they identified strengthen our argument that 8-Ks *could* have an impact on prices. 8-Ks (current material event reports) are much timelier than 10-Ks (annual reports) and 10-Qs (quarterly reports), so it is reasonable to assume that they could also have an impact on market response. Also, the Griffin study was published before the 2004 rule changes, when 8-Ks were even less timely than they are nowadays. Having reviewed the literature, we eventually decided to use a three-day window as we wanted to eliminate the likelihood of confounding events which could occur in a longer window and because a three-day window is consistent with several research studies (e.g. Ball and Kothari 1991; Francis et al 2002; Henry 2008).

We used this initial dataset for a number of purposes. Firstly, we used it to manually analyse the language of 8-Ks and to develop the grammar of financial event phrases (see Section 4.4). We also carried out preliminary experiments with this dataset, using C4.5, an inverted indexing system (IIS), and the most frequently-occurring five-word compound phrases, in an attempt to classify disclosures by likely share price response (Slattery et al 2002).

For the purposes of all our classification experiments, we used only disclosures that had an increase or decrease in share price around the filing date (hereafter referred to as the ‘ups’ and ‘downs’). We also randomly selected 226 ‘downs’ out of the 256 that were downloaded, to ensure an even number of ‘ups’ and ‘downs’. Table 4.4 outlines the total number of newlines, words, and size of the data used in the 7372 dataset. Note how there were more words in the ‘up’ disclosures, which correlates with findings by Hildebrandt and Snyder (1981), Kohut and Segars (1992), and Kroha and Baeza-Yates (2004).

<i>Annret direction</i>	<i># Disclosures</i>	<i># Newlines</i>	<i># Words</i>	<i>Size of dataset (bytes)</i>
<i>Negative (down)</i>	226	361,262	2,576,337	19,590,795
<i>Positive (up)</i>	226	367,825	2,634,903	19,862,833
<i>Total</i>	452	729,087	5,211,240	39,453,628

Table 4.4: Characteristics of data used in the 7372 dataset (number of newlines, words, and size).

### 4.3.2 The First S&P500 Dataset

As our initial dataset only related to specific types of companies (see Section 4.3.1), we wanted to obtain a new listing of 8-Ks for different types of companies and their corresponding closing prices. We contacted several companies requesting data, including Sagentworks, Thomson Reuters, Compustat, PR newswire, and CRSP, but none were able to provide the data we needed, within our budget. We also considered using online systems such as EDGAR Access which enables users to track up to 25 companies on a personalised watch list; however, unlimited access incurs a fee. Whilst some of these systems were affordable, they imposed limitations regarding the number of filings that could be accessed each month and others could not provide the data in the format we needed. We had access to some intraday data via Thomson Reuters' Datastream, but not to closing prices that could be easily correlated with the filing dates in EDGAR indices. For those reasons, we decided to create two new datasets, the first of which will be discussed in the remainder of this section.

As mentioned previously, we randomly chose 50 companies from the S&P 500 index for our two new datasets. This index includes 500 leading companies, all of which are chosen by an S&P Index Committee for their market size, liquidity, and sector representation, amongst other characteristics. All US common equities listed on the NYSE and NASDAQ are eligible for inclusion in the index. This index represents a cross-section of US industry and is considered a leading indicator of US equities (S&P 2012). The initial random listing of companies is presented in Table 4.5.

<b>Ticker</b>	<b>Company name</b>	<b>CIK code</b>	<b>Type of industry</b>
AES	AES Corporation	0000874761	Utilities
AFL	AFLAC Inc.	0000004977	Financials
ALTR	Altera Corp.	0000768251	Information Technology
AMAT	Applied Materials	0000006951	Information Technology
BA	Boeing Company	0000012927	Industrials
BEN	Franklin Resources	0000038777	Financials
BHI	Baker Hughes	0000808362	Energy
CAH	Cardinal Health Inc.	0000721371	Health Care
COG	Cabot Oil & Gas	0000858470	Energy
CVX	Chevron Corp.	0000093410	Energy
DE	Deere & Co.	0000315189	Industrials
DF	Dean Foods	0000931336	Consumer Staples
DGX	Quest Diagnostics	0001022079	Health Care
DHI	D. R. Horton	0000882184	Consumer Discretionary
DUK	Duke Energy	0001326160	Utilities
EMN	Eastman Chemical	0000915389	Materials
EQR	Equity Residential	0000906107	Financials
ESRX	Express Scripts	0000885721	Health Care
GE	General Electric	0000040545	Industrials
HAL	Halliburton Co.	0000045012	Energy
HBAN	Huntington Bancshares	0000049196	Financials
HSP	Hospira Inc.	0001274057	Health Care
IRM	Iron Mountain Incorporated	0001020569	Industrials
KFT	Kraft Foods Inc-A	0001103982	Consumer Staples
KLAC	KLA-Tencor Corp.	0000319201	Information Technology
LNC	Lincoln National	0000059558	Financials
MFE	McAfee	0000890801	Information Technology
MMM	3M Company	0000066740	Industrials
MOLX	Molex Inc.	0000067472	Information Technology
MRO	Marathon Oil Corp.	0000101778	Energy
MS	Morgan Stanley	0000895421	Financials
NKE	NIKE Inc.	0000320187	Consumer Discretionary
NRG	NRG Energy	0001013871	Utilities
NVDA	Nvidia Corporation	0001045810	Information Technology
OKE	ONEOK	0001039684	Utilities
PBI	Pitney-Bowes	0000078814	Industrials
PG	Procter & Gamble	0000080424	Consumer Staples
PLL	Pall Corp.	0000075829	Industrials
PWR	Quanta Services Inc.	0001050915	Industrials
PXD	Pioneer Natural Resources	0001038357	Energy
R	Ryder System	0000085961	Industrials
RHT	Red Hat Inc.	0001087423	Information Technology
RL	Polo Ralph Lauren Corp.	0001037038	Consumer Discretionary
SIAL	Sigma-Aldrich	0000090185	Materials
SLE	Sara Lee Corp.	0000023666	Consumer Staples
STJ	St Jude Medical	0000203077	Health Care
SUN	Sunoco Inc.	0000095304	Energy
TSS	Total System Services	0000721683	Information Technology
TWC	Time Warner Cable Inc.	0001377013	Consumer Discretionary

Table 4.5: Initial random listing of our 50 S&P 500 companies.

We chose the period 1997-2000 for two main reasons. Firstly, by May 1996 all public domestic companies were required to file their disclosures on EDGAR. Secondly, this period was before the events of September 11 2001 and the SEC rule changes which started to come into effect in 2004. One of the objectives of this project was to compare this four-year period with the 2005-2008 period, to see if the 2004 rule changes had any impact on the predictability of closing prices when using financial events in 8-Ks.

Whilst it was easy to access a listing of all 8-Ks filed within a specific time period using the EDGAR indices, the closing prices had to be obtained another way. Unfortunately, EDGAR uses CIKs to identify companies but most online financial websites use ticker symbols, so we had to correlate these before we could link prices to specific filings. Tetlock et al (2008) also encountered similar difficulties matching company details to financial information. To obtain the new dataset, associate it with the closing prices, and prepare it for analysis and classification, it was necessary to do the following:

1. Download the indices for each quarter of each year (\*.idx files). The indices we used were sorted by form type.
2. Merge the four indices into one Excel spreadsheet.
3. Extract the 8-Ks and eliminate all other form types (e.g. 10-Ks). Using these indices, it was possible to determine how many 8-Ks were filed each year. On average, 27,401 8-Ks were filed each year between 1997 and 2000.
4. Use the current S&P 500 ticker symbol and the Edgar 'company search' facility to identify the relevant CIK for each of the randomly-chosen companies.
5. Use a CIK filter in Excel to locate the 8-Ks for the 50 companies. In several cases, the company name and/ or ticker symbol had changed since the filing was made, oftentimes due to mergers and acquisitions. By using the CIK, which is unique to every filer or company, we ensured that the same core filer or company was being used. Even in instances where the current company name was still identical to the original company name, we used the CIK to search for 8-Ks rather than the name, to be consistent.

6. Append the start of the EDGAR URL to the partial filename provided in the form indices ( e.g. append <http://www.sec.gov/Archives/> to ed gar/data/38777/0000038777-99-000462.txt).
7. Run a program which used the ticker symbol and EDGAR filing date to locate the closing prices for days t-1, t, and t+1, on <http://finance.yahoo.com>. We populated a comma-delimited spreadsheet with these prices (.csv file). We encountered a number of issues at this stage. Firstly, the prices program could not deal with filings made during Bank Holiday/ Public Holiday weekends, although it could deal with filings on Saturdays and Sundays of normal weeks. It returned prices of '0' for the former so we had to remove these 8-Ks from the dataset. We also discovered that no prices were available for two of the companies listed in Table 4.5. The first prices issue arose with Dean Foods (DF) which had previously been acquired by Suiza Foods (SZA) but kept the Dean Foods name. It also continued to use the DF ticker symbol after the acquisition. When we searched for historical price data for DF, we could not find data for the filing date in question. When we searched for SZA (the company that acquired DF), no prices were found for any dates, so we had to find a new company to replace it. With regards the second company, no prices were available prior to 2003 for NRG Energy (NRG) so we could not use that company either. The two companies that were randomly chosen to replace DF and NRG were CMS Energy Group (CMS) and O'Reilly Automotive (ORLY).
8. Calculate the annret using the formula  $((Pt+1)/(Pt-1))-1$ , where P stands for the closing share price.
9. Sort 8-Ks by annret. The annret ranged from -0.49 to +0.17 for the revised listing of companies, which had 344 'downs', 102 'no changes', and 357 'ups' between 1997 and 2000.<sup>34</sup>
10. Download the 'ups' and 'downs' only.
11. Filter out disclosures greater than 50kb in size. When we examined the 7372 dataset on a previous occasion, we found that 65% of the 'downs' and 58% of the 'ups' were less than or equal to 50kb in size, equating to an average of 62%. Also, as mentioned previously, Keane (2010) stated that the average submission

---

<sup>34</sup> These figures refer to the *final* listing of S&P 500 companies discussed in the next section. We had to revise the listing of companies as some companies did not have data for both time periods.

on EDGAR is 115kb in size but this figure includes all types of disclosures including 10-Ks. These latter documents, which are also known as the annual reports, are almost always very lengthy. 8-Ks, on the other hand, are supposed to bring recent material events to the attention of investors and the majority of them redirect the reader to other sources of detailed information such as financial statements, press releases, and 10-Ks. We believe that 8-Ks greater than 50kb in size introduce significant noise into the event recognition process, as they include references to previous events and other relatively immaterial events. See Table 4.6 for a breakdown of ‘ups’ and ‘downs’ which were less than or equal to 50kb and for which valid prices were available, in the first S&P dataset. The 548 (256 and 292) disclosures comprised 78% of all the ‘downs’ and ‘ups’ with valid prices, for these 50 companies.

<i>Annret direction</i>	<i># Disclosures</i>	<i># Newlines</i>	<i># Words</i>	<i>Size of dataset (bytes)</i>
<i>Negative (down)</i>	256	83,512	351,890	3,402,519
<i>Positive (up)</i>	292	104,216	455,730	4,255,160
<i>Total</i>	548	187,728	807,620	7,657,679

Table 4.6: Characteristics of data used in the first S&P dataset (number of newlines, words, and size).

- Analyse and classify the 8-Ks less than or equal to 50kb for which prices are available and which have a positive or negative change in annret around the filing date. See Section 4.4 for a discussion of how we developed our list of financial event phrases (FEPs) and Chapter 5 for a discussion of the FEP recognition process. See Chapter 6 for a discussion of our classification experiments.

### 4.3.3 The Second S&P500 Dataset

As mentioned in Section 4.3.2, we wanted to compare the 1997-2000 period with the 2005-2008 period, to see if the 2004 SEC rule changes had any impact on the predictability of closing prices when using the financial events in 8-Ks. Figure 4.1 shows a marked increase in the number of 8-Ks filed each year. If we compare the total for 1997 (24,098) with the total for 2005 (116,282)—the year after the changes

were adopted—we see an almost five-fold increase in the number of filings. This does not come as much of a surprise, however, as the 2004 changes required companies to file more events and within a shorter timeframe (see Section 4.2 for a discussion). It is likely that companies were particularly concerned about adhering to the new rules and decided to err on the side of caution when disclosing events, so as not to leave out anything that might be ‘material’. Interestingly, we should also note a slight decline in filings after 2005, possibly caused by a greater overall understanding of the requirements of the new rules i.e. what they legally had to file versus what they were encouraged to file.

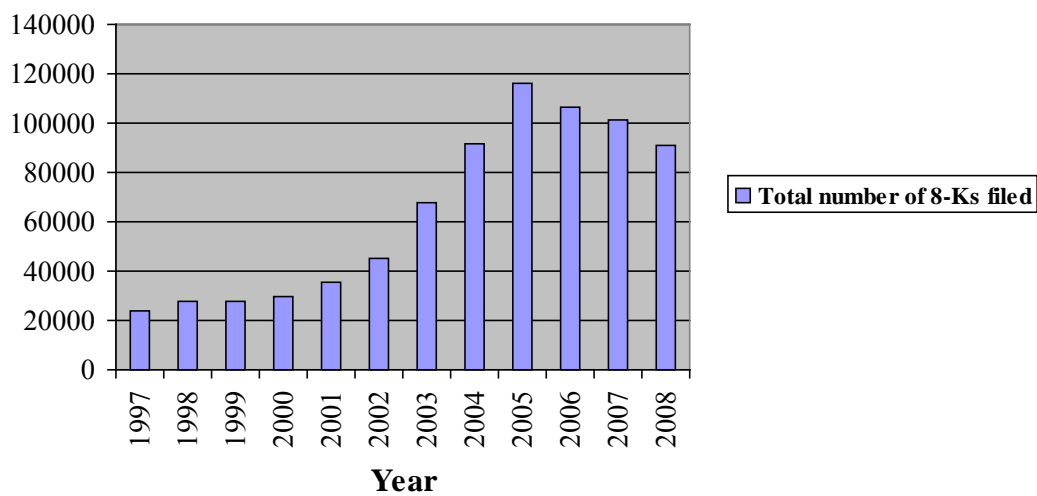


Figure 4.1: Total number of 8-Ks filed each year: 1997 to 2008.

When we used the procedure for obtaining the 8-Ks and the corresponding closing prices that was outlined in Section 4.3.2, we discovered additional problems with the companies that had been randomly chosen. Apart from the issues with SZA and NRG, which had already been rectified by replacing them with two new companies (see Section 4.3.2), we also discovered that no 8-Ks were found for two of the 50 companies for the 2005-2008 period also—Duke Energy (DUK) and Viacom Inc (VIA). A closer analysis revealed that the CIKs for DUK and VIA appeared to change sometime during the two periods due to mergers and acquisitions. During the process of identifying the correct CIK to use for both companies, we also discovered that 8-Ks were missing for six of the 50 companies in the 1997-2000 period, so we had to randomly choose six additional companies to replace them.

There was also one further issue. One company—McAfee (MFE)—was removed from the S&P 500 index in March 2011 so we had to remove that company also and replace it with a new company. Table 4.7 highlights the issues that were identified with the seven companies and the solutions adopted.

<i>Original company (ticker)</i>	<i>Issue (1997-2000)</i>	<i>Issue (2005-2008)</i>	<i>Solution</i>
Duke Energy (DUK)	No 8-Ks found – inconsistent CIKs due to company structural changes	No 8-Ks found – inconsistent CIKs due to company structural changes	Replaced with a new company – Campbell Soup Co (CPB) – in both datasets
Time Warner Cable Inc. (TWC)	No 8-Ks found – inconsistent CIKs due to company structural changes	No 8-Ks found – inconsistent CIKs due to company structural changes	Replaced with a new company – Humana Inc. (HUM) – in both datasets
Viacom Inc (VIA)	No 8-Ks found – inconsistent CIKs due to company structural changes	No 8-Ks found – inconsistent CIKs due to company structural changes	Replaced with a new company – Teradyne Inc. (TER) – in both datasets
Molex Inc (MOLX)	No 8-Ks found - only on EDGAR since 2002	None	Replaced with a new company – Lockheed Martin Corp. (LMT) – in both datasets
Kraft Foods Inc-A (KFT)	No 8-Ks found - only on EDGAR since 2001	None	Replaced with a new company – Hormel Foods Corp /DE/ (HRL) – in both datasets
Hospira Inc. (HSP)	No 8-Ks found - only on EDGAR since 2004	None	Replaced with a new company – Varian Medical Systems <sup>35</sup> (VAR) – in both datasets
McAfee (MFE)	Removed from S&P 500 index in March 2011	Removed from S&P 500 index in March 2011	Replaced with a new company – Fortune Brands Inc. (FO) <sup>36</sup> – in both datasets

Table 4.7: Issues and solutions for the revised listing of S&P companies.

There was also an issue with one of the new companies which subtly changed its name to incorporate a comma. ‘Teradyne Inc.’ changed to ‘Teradyne, Inc.’ midway in the EDGAR indices and this affected the download of prices to our comma separated value (.csv) file. Once the comma was manually removed from the csv file,

<sup>35</sup> Also known as Varian Associates Inc. /DE/ (same CIK code).

<sup>36</sup> Also known as American Brands Inc. /DE/ (same CIK code).

the program successfully downloaded the prices for Teradyne Inc.

When we applied the 50kb size filter to these disclosures, there were 574 ‘downs’ and 672 ‘ups’ remaining with valid prices. It is worth noting here that our 50kb filter gave us access to 78% of the disclosures in the 1997-2000 dataset whereas it only gave us access to half (49%) of the available ‘ups’ and ‘downs’ in the 2005-2008 dataset (see Section 4.3.2). This is indicative of the changes that came about in 2004 and in subsequent years, which required greater disclosure of material events and probably lengthier 8-Ks as a result. The annrets for this final dataset ranged from -0.58 to +0.50. Table 4.8 outlines the characteristics of the data used in the second S&P dataset.

<i>Annret direction</i>	<i># Disclosures</i>	<i># Newlines</i>	<i># Words</i>	<i>Size of dataset (bytes)</i>
<i>Negative (down)</i>	574	292,980	1,234,809	13,941,978
<i>Positive (up)</i>	672	338,401	1,447,346	16,491,343
<i>Total</i>	1,246	631,381	2,682,155	30,433,321

Table 4.8: Characteristics of data used in the second S&P dataset (number of newlines, words, and size).

Our final listing of S&P companies is provided in Table 4.9, with new companies in bold type near the end of the table.

<b>Ticker</b>	<b>Company name</b>	<b>CIK code</b>	<b>Type of industry</b>
AES	AES Corporation	0000874761	Utilities
AFL	AFLAC Inc.	0000004977	Financials
AMAT	Applied Materials	0000006951	Information Technology
BA	Boeing Company	0000012927	Industrials
SLE	Sara Lee Corp.	0000023666	Consumer Staples
BEN	Franklin Resources	0000038777	Financials
GE	General Electric	0000040545	Industrials
HAL	Halliburton Co.	0000045012	Energy
HBAN	Huntington Bancshares	0000049196	Financials
LNC	Lincoln National	0000059558	Financials
MMM	3M Company	0000066740	Industrials
PLL	Pall Corp.	0000075829	Industrials
PBI	Pitney-Bowes	0000078814	Industrials
PG	Procter & Gamble	0000080424	Consumer Staples
R	Ryder System	0000085961	Industrials
SIAL	Sigma-Aldrich	0000090185	Materials
CVX	Chevron Corp.	0000093410	Energy
SUN	Sunoco Inc.	0000095304	Energy
MRO	Marathon Oil Corp.	0000101778	Energy
STJ	St Jude Medical	0000203077	Health Care
DE	Deere & Co.	0000315189	Industrials
KLAC	KLA-Tencor Corp.	0000319201	Information Technology
NKE	NIKE Inc.	0000320187	Consumer Discretionary
CAH	Cardinal Health Inc.	0000721371	Health Care
TSS	Total System Services	0000721683	Information Technology
ALTR	Altera Corp.	0000768251	Information Technology
BHI	Baker Hughes	0000808362	Energy
CMS	CMS Energy	0000811156	Utilities
COG	Cabot Oil & Gas	0000858470	Energy
DHI	D. R. Horton	0000882184	Consumer Discretionary
ESRX	Express Scripts	0000885721	Health Care
MS	Morgan Stanley	0000895421	Financials
ORLY	O'Reilly Automotive	0000898173	Consumer Discretionary
EQR	Equity Residential	0000906107	Financials
EMN	Eastman Chemical	0000915389	Materials
IRM	Iron Mountain Incorporated	0001020569	Industrials
DGX	Quest Diagnostics	0001022079	Health Care
RL	Polo Ralph Lauren Corp.	0001037038	Consumer Discretionary
PXD	Pioneer Natural Resources	0001038357	Energy
OKE	ONEOK	0001039684	Utilities
NVDA	Nvidia Corporation	0001045810	Information Technology
PWR	Quanta Services Inc.	0001050915	Industrials
RHT	Red Hat Inc.	0001087423	Information Technology
<b>CPB</b>	<b>Campbell Soup</b>	<b>0000016732</b>	<b>Consumer Staples</b>
<b>HUM</b>	<b>Humana Inc.</b>	<b>0000049071</b>	<b>Health Care</b>
<b>TER</b>	<b>Teradyne Inc.</b>	<b>0000097210</b>	<b>Information Technology</b>
<b>LMT</b>	<b>Lockheed Martin Corp.</b>	<b>0000936468</b>	<b>Industrials</b>
<b>HRL</b>	<b>Hormel Foods Corp.</b>	<b>0000048465</b>	<b>Consumer Staples</b>
<b>VAR</b>	<b>Varian Medical Systems</b>	<b>0000203527</b>	<b>Health Care</b>
<b>FO</b>	<b>Fortune Brands Inc.</b>	<b>0000789073</b>	<b>Consumer Discretionary</b>

Table 4.9: Final random listing of S&P 500 companies.

## 4.4 The Causal Effect of 8-K Disclosures in Various Windows

In Section 4.3.1, we discussed why we chose a three-day window using days  $t \pm 1$ , where day  $t$  is the 8-K filing date. In our discussion, we referred to various studies that used different sized windows. In this section, we examine the causal effect of 8-Ks on prices, by considering the wider price dynamics in which 8-Ks are filed. Specifically, for all 8-Ks in the first and second S&P datasets, for which prices were available, we examine the share price changes using three window sizes: days  $t \pm 1$ ,  $t \pm 5$ , and  $t \pm 100$ . Table 4.10 shows the range of share price changes for the three window sizes.

	<i># Disclosures (Range of share price changes using <math>t \pm 1</math>)</i>	<i># Disclosures (Range of share price changes using <math>t \pm 5</math>)</i>	<i># Disclosures (Range of share price changes using <math>t \pm 100</math>)</i>
<i>First S&amp;P Dataset (1997-2000)</i>	803 (-0.49 to 0.17)	803 (-0.68 to 0.47)	778 (-0.89 to 2.07)
<i>Second S&amp;P Dataset (2005-2008)</i>	3247 (-0.58 to 0.50)	3247 (-0.63 to 0.71)	3226 (-0.90 to 1.39)

Table 4.10: Range of share price changes for various windows (both datasets, available prices)

Whilst the ranges of share price changes were greater for the second dataset for two of the three window sizes, compared to the first, we should bear in mind that there were significantly more disclosures in the second dataset. Regardless of dataset, we can also see that the range was greater as we moved from days  $t \pm 1$ , to  $t \pm 5$ , to  $t \pm 100$ , indicating greater overall fluctuations in the market for larger window sizes.

To further examine market reactions, we generated histograms and box plots to summarise the share price changes for both datasets, for the three window sizes. As can be seen in Appendices 2, 4, 6, 8, 10, and 12, the box plots show little variation around 0, regardless of which window we use, although some box plots are slightly negatively/ positively skewed. Table 4.11 summarises the price changes for 90% of the data. For example, in the second S&P dataset, using the  $t \pm 1$  window, 90% of the

data lies in the range of  $0\pm 0.07$ .

	<i>Majority price changes using <math>t\pm 1</math></i>	<i>Majority price changes using <math>t\pm 5</math></i>	<i>Majority price changes using <math>t\pm 100</math></i>
<i>First S&amp;P Dataset (1997-2000)</i>	$0\pm 0.10^{37}$	$0\pm 0.15$	$0\pm 0.60$
<i>Second S&amp;P Dataset (2005-2008)</i>	<b><math>0\pm 0.07</math></b>	$0\pm 0.12$	$0\pm 0.50$

Table 4.11: 90% of data lies within these ranges (both datasets, available prices).

Whilst the histograms and box plots show that all the data is pretty much centred at 0, we did find that the curve was wider for larger windows (compare, for example, the width of the boxplots in Appendices 2 and 4, both of which use the same scale). This highlights the larger range of price changes over the larger window, presumably caused by various factors affecting prices, in the larger window.

We then examined some extreme outliers (identified by asterisks in the boxplots in Appendices 2, 4, 6, 8, 10, and 12), to see if we could identify a link between the extreme change in price for those companies and the release of 8-Ks. We generated time series plots for eight of these outliers, using closing prices over a twelve-month period (with the 8-K filing date roughly in the middle). We took two extreme negative outliers and two extreme positive outliers from both datasets (1997 to 2000 and 2005 to 2008) and examined the 8-K content to see if it could possibly have caused the extreme change, also bearing in mind the trends before and after the 8-K filing.

In Appendix 13, we can see the time series for Halliburton Co, which released a filing on 22<sup>nd</sup> July 1997, according to the EDGAR indices. The filing date stated in the 8-K is 23<sup>rd</sup> July. On this occasion, the 8-K discussed an *earlier* offer, sale, and delivery of ‘notes’ and not the two-for-one stock split, which was actually the cause of the

---

<sup>37</sup> Price changes are rounded to two decimal places.

extreme drop in closing price. The stock split reduced the closing price from 84.48 on 21<sup>st</sup> July to 43.66 on 22<sup>nd</sup> July. Whilst we can assume that the stock split was discussed in an earlier 8-K, we cannot link the content of the 22<sup>nd</sup> July filing with the change in closing price. Ignoring the stock split, the closing prices are reasonably stable otherwise, before and after the 8-K filing date.

The second extreme outlier we examined also concerned a stock split. Appendix 14 relates to Morgan Stanley Dean Witter & Co. which filed an 8-K on 26<sup>th</sup> January 2000, according to the indices. However, the official filing date stated on the 8-K is 3<sup>rd</sup> February, several days later. A two-for-one stock split, which was previously announced on 20<sup>th</sup> December 1999, took place the day of the filing (26<sup>th</sup> January) and the filing discussed the details of the stock split. Presumably, the market would have already reacted to the impending split by the time the stock split took effect, as the company disclosed this event a month before. The stock split caused the closing price to be adjusted from 132.88 on 26<sup>th</sup> January to 67.56 the following day. Whilst we cannot say that the *discussion* of the stock split in the 8-K filed on 26<sup>th</sup> January caused investors to react in a significant way, we can say that the closing price was adjusted *because of* the stock split, which was discussed in this 8-K. As mentioned earlier, the stock split was *first* announced on 20<sup>th</sup> December 1999 and we can observe a slight rise in closing price the day after the stock split was announced. However, looking at the general trend for this company, the closing prices are somewhat unstable anyway, so we cannot be certain that the initial disclosure about the stock split had any impact on price.

In Appendix 15, we can see the time series for Nike Inc, which filed an 8-K on 18<sup>th</sup> September 1998, according to the indices. The date stated on the 8-K is 21<sup>st</sup> September, a few days later. In the 8-K, Nike discussed declining quarter results and workforce reductions, which were previously disclosed on 17<sup>th</sup> September (the day before the 8-K), in a separate press release. As the news was first disclosed on 17<sup>th</sup> September, we would expect the negative news (particularly the declining quarter results) to have an impact later that day, or on the days that followed. However, the closing price started to increase on 17<sup>th</sup> September (33.22) and continued to climb until 21<sup>st</sup> September (38.45); the latter date being the date stated on the filing itself.

There was then a decline in price to 37.96 (22<sup>nd</sup> September). Looking at the overall trend for this company, we can see that it is fairly stable overall and it seemed to be recovering from the fairly significant decline that occurred between July and September. As the 8-K contained negative news, we would not have expected the closing price to increase in the days that followed, so we cannot infer that this 8-K had any impact on the price.

In Appendix 16, we can see the time series for Red Hat Inc., which filed on 16<sup>th</sup> November 1999. The date stated on the 8-K was 17<sup>th</sup> September. In this example, the 8-K referred to a press release, which was filed on 15<sup>th</sup> November. In the press release, the company announced a merger. The press release was incorporated into the 8-K also. On 15<sup>th</sup> November, the closing price was 105.37 and in the days that followed, the price increased<sup>38</sup>. We could infer that this increase was caused by the merger discussed in the press release and 8-K; it is not possible, however, to determine which news source, if any, actually caused this price change. However, on 19<sup>th</sup> November the price took a slight dip but then steadily increased from 22<sup>nd</sup> (124.5) to 29<sup>th</sup> November (236.63). Looking at the bigger picture, the merger was due to be closed by 27<sup>th</sup> January 2000, so it is *possible* that the instability that followed (see end of November 1999 to January 2000 in Figure 4.4) was caused by company restructuring and other uncertainties caused by the merger.

As mentioned earlier, we also examined four extreme outliers from the 2005 to 2008 dataset. As two of these outliers belonged to the same company, we only generated one time series plot (rather than two). Appendix 17 shows the time series for Lincoln National Corp. which filed an 8-K on 10<sup>th</sup> October 2008. This date matches the date stated within the 8-K. In the 8-K, they referred to a press release that was filed *the same day* announcing preliminary financial results. In general, the financial results were promising but they did announce that they were going to reduce dividends and that they would be suspending the share repurchase agreement for the rest of the year. The day before the filing (9<sup>th</sup> October), the closing price was at its lowest point in at least six months (18.31) but on the day of the filing, it increased (to 23.95) and it

---

<sup>38</sup> For this company, historical prices were only available back to 11<sup>th</sup> August 1999, so Figure 4.4 does not cover a twelve-month period.

continued to rise until 14<sup>th</sup> October (30.36). We could *infer* that this brief improvement in performance was caused by the news that was simultaneously released in a press release and 8-K. We cannot, however, determine which source, if any, caused the price change. The general trend for this company shows that the closing prices are not particularly stable.

Lincoln National Corp also filed an 8-K on 19<sup>th</sup> November 2008, according to the indices. As discussed in Section 4.2, after 2004, companies were required to file 8-Ks in a timelier manner; in this case, the date stated on the 8-K matches the official filing date in the indices. In the 8-K, the company discussed a conference for investors and managers that took place that same day. In the 8-K, they elaborated on the definitions of the non-GAAP measures discussed during the conference. Looking again at Appendix 17 more closely, we can see that the closing price was in general decline since the 19<sup>th</sup> of September. The day after the filing (20<sup>th</sup> November), the closing price reached its lowest point in twelve months (looking at six months on either side of the filing date). We could infer that the information provided during the investor's conference *and* in the 8-K helped to clarify any concerns about the company's financial position and this led to the slight improvement in performance in the two-three months that followed. However, we cannot be certain that the improvement was caused by the conference or the 8-K; it is also possible that the market had "bottomed-out".

In Appendix 18, we can see the time series for O'Reilly Automotive Inc., which filed on 16<sup>th</sup> June 2005. The date on the 8-K matches the official filing date in the indices. In this 8-K, the company announced that it would be presenting at a 'growth stock conference' a few days later. On this occasion, the extreme change in closing price was not caused by the content of the 8-K but by an event that was presumably disclosed in an earlier 8-K—a two-for-one stock split, which caused the closing price to be adjusted from 58.99 (15<sup>th</sup> June) to 29.78 (16<sup>th</sup> June). Apart from the stock split, the general trend shows that the closing prices were reasonably stable.

The time series for Dr Horton Inc. is provided in Appendix 19. The index and 8-K filing dates match (25<sup>th</sup> November 2008). The 8-K refers to a press release of the

*same day*, which announced the results for the quarter. Whilst the results were mainly negative, they did declare that they would be issuing a dividend. This news follows a period of steep decline during October. Whilst there was a brief increase in price the day after the filing (from 6.9 to 7.52), the price dropped a gain until 1<sup>st</sup> December (to 6.02). Despite various fluctuations, the overall trend after the 8-K is a rising one. Whilst we could infer that the price increase the day after the 8-K was filed was caused by the content of the 8-K, we believe that the news is more negative than positive and this change in price may instead have been caused by random fluctuations.

In a number of cases, we can infer from our subjective analysis that the change in closing price around the filing date was caused by the content of the 8-Ks but we cannot deduce there is a definite causal link. In the newer dataset, for example, two of the 8-Ks were released the same day as press releases describing the same event, so it is not possible to deduce which, if indeed any, of these sources actually caused the price change. In some cases, the time series were quite unstable anyway, so it is possible that these fluctuations would have occurred even if an 8-K had not been filed. The most we can conclude from our analysis is that 90% of the price changes lie in the range of  $0 \pm 0.10$  (first dataset) and  $0 \pm 0.07$  (second dataset) when using a days  $t \pm 1$  window and that the range is larger for larger windows (days  $t \pm 5$  and  $t \pm 100$ ) (see Table 4.11). As one might expect, prices fluctuate more frequently in larger windows, and often to a greater extent, as a greater number of “events” have taken place during that time. As there was so much variation in the larger windows, it is less likely that the price changes we observed were caused by *one specific event* (e.g. the release of an 8-K). Finally, taking the analysis of our three-day window (days  $t \pm 1$ ) one step further, we found that almost 84% (83.9%) of the data in the second S&P dataset had price changes equal to or greater than  $0 \pm 0.01$  (or  $\geq \pm 1\%$  change in price). We maintain that a 1% (or greater) change in price within a three-day window on either side of an 8-K filing date is worthy of examination and propose that these changes may be caused by events discussed in the disclosures.

## 4.5 Identification of Financial Event Phrases (FEPs)

We manually analysed the content of 185 disclosures, comprising the top 56 ‘ups’, top 56 ‘downs’, and all 73 ‘no changes’ in the initial dataset (see Section 4.3.1). By ‘top’, we mean the disclosures with the greatest changes in annret. We chose these disclosures because we felt they were more likely to contain content that could impact the share price. During the manual analysis, we identified 20 major event categories and several key event types within those categories. See Table 4.12 for the initial ontology.

<i>Major event category</i>	<i>Key event types</i>
1. Accountant dismissal and appointment	1.1 Accountant dismissal 1.2 Accountant appointment
2. New/ resigned/ dismissed employee(s)	2.1 New employees/promotions/election 2.2 Potential new employees 2.3 Remain as employees 2.4 Resign/leave 2.5 Dismiss/layoff employees 2.5 Potential problems with employees
3. Agreement and plan of merger	3.1 Acquisition agreement and plan of merger 3.2 Amendment to acquisition agreement and plan of merger
4. Agreement and plan of reorganization	4.1 Agreement and plan of reorganization 4.2 Close agreement and plan of reorganization
5. Securities purchase agreement	5.1 Securities purchase agreement
6. Stock option agreement	6.1 Stock option agreement
7. Rights plan	7.1 Rights plan
8. Indenture	8.1 Indenture
9. Acquisition of stock	9.1 Acquisition of stock
10. Sale of stock	10.1 Sale of stock
11. Purchase and sale of stock	11.1 Purchase and sale of stock
12. Private placement	12.1 Private placement
13. Issue of stock	13.1 Issue of stock
14. Goodwill	14.1 (Has) goodwill 14.2 No goodwill
15. Loss	15.1 Loss
16. Claim of copyright infringement	16.1 Claim of copyright infringement
17. Counterclaim	17.1 Counterclaim
18. Granted and exception	18.1 Granted an exception
19. Change(d) state of incorporation	19.1 Change(d) state of incorporation
20. Material adverse effect	20.1 Could have a material adverse effect 20.2 Shall not have any/ no material adverse effect

Table 4.12: Initial ontology from the manual analysis of disclosures.

We then identified all the disclosures that contained keywords relating to each major event category (see Table 4.12 for the list of categories), using a combination of search and concordance tools. Sample search keywords included *dismissal*,

*appointment, resign, and merger*. Using search tools, we manually analysed disclosures which matched the search keywords and extracted the key phrases that described the relevant events. Using concordance tools, we manually analysed all the returned sentences that matched the search keywords and extracted the most critical part of the sentence. These phrases were edited manually to remove references to company names, product names, and financial values. Sample phrases included *consummated a private offering, reported record quarterly net income, and entered into an agreement to acquire all of the outstanding capital stock*. This filtering created a ‘best choices’ list of financial event phrases (FEPs) for each event (see also Appendix 20 for a selection of FEPs for employment-related events).

We then fine-tuned several key event types. For example, some new events were identified (e.g. ‘amended by-laws’ and ‘product offering’), some events were split (e.g. ‘rights plan’ was split into ‘amended rights plan/ issue/ offer/ agreement’ and ‘rights plan/ issue/ offer/ agreement’), and others were merged (e.g. ‘agreement and plan of merger’ was merged with ‘agreement and plan of reorganization’ to form ‘acquisition agreement and plan of merger/ reorganization’). See Table 4.13 for our final ontology of FEPs. We developed our own ontology rather than use the official ontology devised by the SEC because the SEC’s ontology was quite limited when we developed ours (pre 2004). Whilst we believe our ontology captures many significant event items, there is scope for further development of the ontology considering the more recent filing requirement changes (see Table 4.3).

<i>Key event categories</i>	<i>Key event types</i>
Accountant dismissals and appointments	1 Accountant appointment 2 Accountant dismissal
New/ resigned/ dismissed/ remaining employee(s) (not related to accountants)	3 Personnel appointments/ promotions 4 Personnel resigning/ departing 5 Potential employment problems 6 Potential new personnel 7 Remain as personnel 8 Layoff/ dismissal of personnel
Acquisitions, mergers and reorganizations	9 Acquisition agreement and plan of merger/ reorganization 10 Close an agreement and plan of merger/ reorganization 11 Amended acquisition agreement and plan of merger / reorganization
Stock-related events	12 Securities purchase agreement 13 Private placement of stock 14 Cancel private placement of stock 15 Public stock offering

	16 Cancel or postpone stock offering 17 Stock option agreement/ plan 18 Stock or note conversion 19 Stock split 20 Reverse stock split 21 Adopt resolution to designate/ create stock
Rights plans	22 Rights plan/ issue/ offer/ agreement 23 Amended rights plan/ issue/ offer/ agreement
Income/ loss/ revenue	24 Will/ has generate(d) income 25 Will/ has generate(d) income decrease 26 Will/ has generate(d) loss 27 Will/ has generate(d) revenue decrease 28 Will/ has generate(d) revenue
Dividend distributions	29 Dividend distribution
Purchase agreements	30 Purchase agreement 31 Amended purchase agreement
Goodwill-related events	32 Allocation/ writeoff/ amortization of goodwill 33 No goodwill
Indentures	34 Indenture
Adoptions of standards or strategies	35 Has/ will adopt financial standard(s) 36 Adopt strategy/ plan
Listing-related events	37 Meets/ expects to meet listing requirements 38 Has not/ might not meet listing requirements 39 Granted an exception
Amendments to reports and laws	40 Has/is/will amend general agreement/ report/ document/ information/ by-laws
Other events	41 Alleged copyright infringement 42 Breach 43 Filed a counterclaim 44 Owes or paid compensation 45 Change state of incorporation 46 Relocation 47 Product offering 48 Could/ would/ can/ expect material adverse effect 49 Could/ would/ cannot expect material adverse effect

Table 4.1 3: Final ontology of financial event phrases ( FEPs) used in the grammar.

We decided to develop our own grammar of financial event phrases, rather than use phrases identified in other studies, because the other studies used online news stories or messages, not disclosures (see, for example, Wüthrich et al 1998, Cho et al 1999, Thomas and Sycara 2000, Das and Chen 2001, Peramunetilleke and Wong 2002, Thomas 2003, Seo et al 2004, van Bunningen 2004, Schumaker and Chen 2006 and Mittermayer and Kolmayer 2006a in Chapter 3). The language in disclosures is quite different to the language used in news headlines, stories and discussion board

postings (see Section 2.3 for a discussion of the language used in financial reports and news). We also decided to identify phrases for as many event types as possible, rather than focus on specific events, to avoid what Fama (1998) referred to as “dredging for anomalies” (p.287). See Section 5.3 for an analysis of how many events were identified in the datasets by our prototype recogniser.

Whilst we were generating the FEP lists, we also generated a list of named entities (NEs). In the Message Understanding Conferences (MUC), named entities were defined as “proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts” (Chinchor 1998, n.p.). We identified NEs such as accountant names (e.g. *KPMG Peat Marwick LLP* and *Price Waterhouse LLP*) and types of employee (e.g. *Senior Vice President, Strategy, Finance and Administration* and *Chairman of the Board of Directors*). We also developed a list of types of financial object (TFO)—for example, we identified 84 different types of loss (e.g. *loss before provision for income tax* and *unrealized loss on investments*), 193 types of stock, right or option (e.g. *redeemable convertible preferred shares* and *preferred stock*) and 72 types of stock option agreement or plan (e.g. *stock option agreement* and *supplemental stock plan*). The prototype FEP recogniser (see Section 5.2) used the NEs and TFOs to ensure that as many possible variations of a financial event would be recognised and that where FEPs could not be recognised in an 8-K, at least NEs and/ or TFOs would be written to the output, to facilitate classification later on. See Appendix 21 for our list of FEP types, NEs, and TFOs.

Finally, we also generated a list of the most frequently-occurring keywords. We then fine-tuned this list by removing “useless” stop words and other words to create a list of 1,568 interesting keywords. Whilst not all these words were of a financial nature, they all appeared frequently in our 8-Ks (see Appendix 22). These words were used as additional features in the classification experiments described in Chapter 6. We did not use the keywords identified by Loughran and McDonald (2011b) as they had not been published at the time of our experiments. Whilst their keywords related to 10-Ks, we believe there could be significant cross-over in language; for that reason, we propose that their keywords be used with ours, in future research (Section 7.4).

## 4.6 Summary

In this chapter, we presented the rationale for choosing 8-K disclosures as the focus of our study. We examined the format of 8-Ks and other relevant regulations, such as filing deadlines. We also described the datasets we used in our research and discussed the wider price dynamics in which our 8-Ks were filed. We found that the majority of 8-Ks (83.9%) in the second S&P 500 dataset had price changes greater than or equal to 1%, within a three-day window around the filing date, and proposed that some of these price changes may have been caused by the financial events described in the 8-Ks. We discussed how we identified the financial event phrases (FEPs) that were subsequently used to recognise FEPs in 8-Ks. In the next chapter, we will discuss how we developed our prototype FEP recogniser. We will also discuss the findings from our automatic pattern analysis of recognised FEPs.

# Chapter 5: Automatic Analysis of Financial Events in 8-K Disclosures

## 5.1 Outline

In the previous chapter, we presented our rationale for choosing Form 8-K disclosures. We examined the format of 8-Ks and summarised other relevant regulations, such as filing deadlines. We also described the datasets we used in our research and discussed the wider price dynamics in which our 8-Ks were filed. We then discussed how we identified the financial event phases (FEPs) that were subsequently used to recognise FEPs in 8-Ks.

In this chapter, we discuss the development of the prototype FEP recogniser. In Section 5.2 we present a brief overview of the techniques we used to recognise FEPs. In Section 5.3, we present an overview of the features of the *recognised* output in both datasets, using automatic pattern analysis techniques. We also examine both datasets together, with a view to identifying possible trends or patterns in the ‘downs’ and ‘ups’. Finally, in Section 5.4, we provide a summary of the chapter.

## 5.2 Recognition of FEPs in Form 8-Ks

In Section 5.2.1, we discuss how we developed our FEP recogniser using a simple and well-proven technique based on a cascade of non-deterministic top-down parsers, implemented as Definite Clause Grammars, and written in Prolog (Pereira and Warren 1980). By non-deterministic we mean the recogniser makes arbitrary choices of which FEP grammar rules to use and then revises those choices later. By top-down we mean the recogniser works from the root down to sentence level, using the FEP rules in the recogniser. In Section 5.2.2, we outline some issues we encountered when developing the recogniser.

### 5.2.1 Development of the FEP Recogniser

As discussed in Section 4.4, we developed a list of FEPs, named entities (NEs), and types of financial object (TFOs), all of which were used together to ensure that as many possible variations of an event would be recognised and that where an FEP could not be recognised, at least a NE and/or TFO would be written to the output file. We will now outline how the FEPs, NEs, and TFOs are recognised. Figure 5.1 outlines all the tasks that need to be carried out to recognise them and this is followed by a textual description of the process.

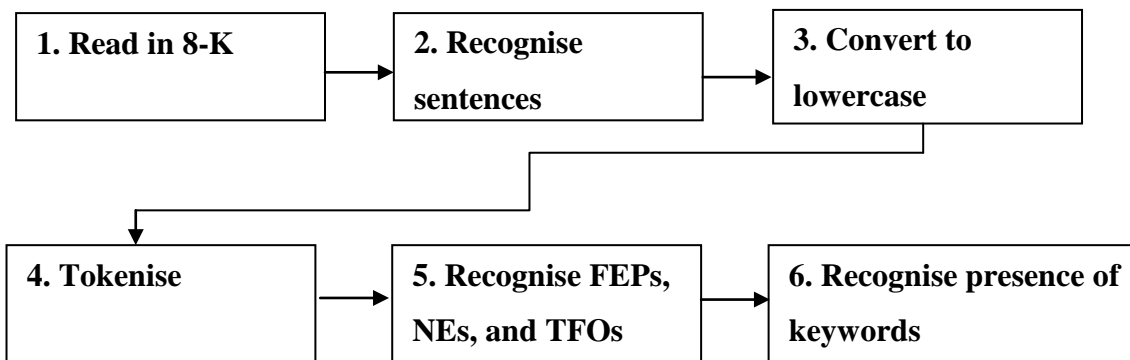


Figure 5.1: The FEP, NE, and TFO recognition process.

1. The full 8-K text is read in as a single string.
2. The 8-K string is split into a list of sentences.
3. Each sentence string is converted to lowercase.
4. Each lowercase sentence string is tokenised. Full-stops within words are retained. Each tokenised sentence is now a list of atoms.
5. The FEPs, NEs, and TFOs within each tokenised sentence are recognised and written to an output file.
6. The full text of the 8-K is compared to a list of 1,568 hand-chosen keywords, to identify which keywords exist. These are also written to an output file.

Once the FEPs, NEs, TFOs, and keywords have been recognised, the output is prepared for C4.5 or SVM-Light classification. See Sections 6.2.1 and 6.3.1 for an overview of both algorithms and the data formats required.

### 5.2.2 Issues with the FEP Recogniser

As mentioned previously, we encountered a number of issues when developing the recogniser. The first issue related to the financial event phrases. It was difficult at times to interpret the meaning of a phrase and to determine if a phrase was one type of event or another. The stock-related events, in particular, were difficult to interpret at times (see Table 4.11 for a listing). We are not yet convinced that this is the ideal categorisation of stock-related events and recommend that further work is carried out in this area (see Chapter 7). Other researchers who highlighted similar kinds of difficulties when trying to interpret the meaning of financial text include Hildebrandt and Snyder (1981), who discussed the importance of context; Thomas (1997), who studied condensations and contractions; and Gillam et al (2002), who discussed problems with negation, double-negation and ambiguous terms (see Section 2.3 for a discussion).

Another related issue was that the lists of FEP phrases, named entities, and types of financial object were initially compiled using 8-Ks filed between 1994 and 1998. These 8-Ks referred to events that were, at the time, considered to be of material interest to investors, so it is possible that the content bore a close relationship with the subsequent share price changes around the filing dates. However, the second S&P 500 dataset referred to 8-Ks filed *after* the 2004 SEC rule changes, which required companies to file many more types of events. It is possible that the newer 8-Ks contain events that cannot be identified by our recogniser. Nonetheless, when identifying the phrases for the recogniser, we did not focus on the nine items that were initially required (see Table 4.1); instead, we searched for all types of events that we believed were of interest to investors and we identified interesting phrases accordingly. Even before the 2004 rule changes, filers were supposed to file ‘material events’, so we believe that our recogniser may not be as restrictive as one might initially assume.

As mentioned in Section 4.3.2, we decided only to parse 8-Ks that were less than or equal to 50kb in size, as an initial analysis of the initial dataset revealed that, on average, 62% of the ‘ups’ and ‘downs’ were less than or equal to 50kb in size. We also felt that larger 8-Ks would be too noisy. However, in the first and second S&P

500 datasets, this filtering of disclosures reduced our datasets by 22% and 51% respectively. In a real-world implementation, it would be better to extract the most significant sections first and then search for items in the remaining content.

Once the recognition process was complete, we had to perform some post-processing, to format the output in a manner that would be suitable for performing pattern analysis (see Section 5.3) and for classifying the likely share price response of disclosures based on their FEPs (see Chapter 6).

### **5.3 Automatic Pattern Analysis of FEPs**

We carried out a preliminary analysis of FEPs found by the recogniser, to see if we could identify any interesting patterns about events in 8-Ks (see Sections 5.3.1 and 5.3.2), prior to carrying out the classification experiments (Chapter 6). We also used these analysis findings to indirectly examine the performance of the recogniser.

#### **5.3.1 Pattern Analysis of FEPs in the First S&P 500 Dataset**

As outlined in Chapter 4, the first S&P 500 dataset relates to 1997 to 2000, which was before the SEC rule changes which required companies to file more events, came into effect. This dataset had 807,620 words and was 7.7MB in size (see Table 4.6).

We first determined the total number of FEPs recognised in all files, including duplicate FEPs (i.e. where the same event was mentioned more than once in the same 8-K). Table 5.1 shows the total number of FEPs recognised in all files, the average number of FEPs recognised per file and the minimum and maximum number of FEPs recognised in any one file. The recogniser identified 274 FEPs in the ‘downs’ and 260 FEPs in the ‘ups’, including duplicates. Note how more FEPs were recognised in the ‘downs’, even though there were less ‘down’ 8-Ks. On average, there was only 1 FEP recognised per file (rounded to the one decimal place), although this varied from 0 to 28 FEPs (‘downs’) and from 0 to 18 (‘ups’).

	1997-2000 'downs' (256 8-Ks)	1997-2000 'ups' (292 8-Ks)
Total number of FEPs found in all files, including duplicates	274	260
Average number of FEPs per file	1.1	0.9
Minimum and maximum number of FEPs found in any one file	0-28	0-18

Table 5.1: Total, average, and range of FEPs recognised in all 'downs' and 'ups', including duplicates (first S&P 500 dataset).

We then determined the number of disclosures for which 0, 1, 2, 3... FEPs were identified by the recogniser (see Table 5.2). The relatively high number of instances of disclosures which had 0 recognised FEP types (coincidentally 69.1% and 69.2% for the 'downs' and 'ups' respectively) highlights how difficult it can be to write a definitive grammar, particularly for the financial domain which has very little regulation in terms of writing style. As discussed in Chapter 2, the language used in financial reports and articles can vary greatly depending on the author, purpose, and topic being discussed (e.g. good or bad news). Despite the fact that our recogniser comprises over 2,000 FEP phrases, it succeeded in identifying FEPs in only 31% of the 'downs' and 31% of the 'ups'. For these remaining 8-Ks (i.e. 79 'downs' and 90 'ups'), we examined the number of FEPs that were recognised in each 8-K. Table 5.2 also shows the total number of FEPs recognised for the remaining 'downs' and 'ups', including duplicate FEPs.

# FEPs identified (incl. duplicates)	# Downs (256 8-Ks)	# Ups (292 8-Ks)	Total # disclosures (548 8-Ks)
0	177 (69.1%)	202 (69.2%)	379 (69.2%)
1	41 (16.0%)	39 (13.4%)	80 (14.6%)
2	8 (3.1%)	20 (6.8%)	28 (5.1%)
3	4 (1.6%)	9 (3.1%)	13 (2.4%)
4	5 (2.0%)	7 (2.4%)	12 (2.2%)
5	5 (2.0%)	1 (0.3%)	6 (1.1%)
6+	16 (6.3%)	14 (4.8%)	30 (5.5%)

Table 5.2: Number of 'downs' and 'ups' which had 0 or more FEP types (first S&P 500 dataset).

As Table 5.2 shows the number of FEPs recognised, including duplicates, we also determined how many *different FEP types* were identified by the recogniser. Table 5.3 shows that 18 of the 49 FEP types were identified in the 79 ‘downs’ and there were 110 unique occurrences across those 79 ‘downs’, when duplicate FEPs within individual 8-Ks were removed. In the 90 ‘ups’, 19 of the 49 FEP types were identified and there were 119 unique occurrences when duplicate FEPs within individual 8-Ks were removed. Appendix 23 lists the different FEP types that were recognised for this dataset.

	<i>Downs</i> (256 8-Ks)	<i>Ups</i> (292 8-Ks)
# Different FEP types found	18/49 (36.7%)	19/49 (38.8%)
# Unique occurrences (no duplicates within 8-Ks)	110	119

Table 5.3: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (first S&P 500 dataset).

A recent study by Lerman and Livnat (2009) found that 78% of the 8-Ks in their dataset were single-event filings. When we only considered the 8-Ks for which FEPs were recognised (see Table 5.2), and unique occurrences of FEPs (Table 5.3), we found that 72.8% of the 8-Ks were single-event filings (see the second row, fourth column in Table 5.4). However, we must remember that our recogniser only succeeded in finding FEPs in 30.9% of the ‘downs’ and 30.8% of the ‘ups’ (see Table 5.2), so this figure could go either way with a more efficient recogniser. In reality, we would expect to find at least one FEP type in each 8-K.

# Unique FEPs identified (no duplicates)	<i>Downs</i> (110 unique FEP occurrences)	<i>Ups</i> (119 unique FEP occurrences)	<i>Downs and Ups</i> (229 unique FEP occurrences)
1	59 (74.7%)	64 (71.1%)	123 (72.8%)
2	12 (15.2%)	23 (25.6%)	35 (20.7%)
3	5 (6.3%)	3 (3.3%)	8 (4.7%)
4	3 (3.8%)	0 (0%)	3 (1.8%)
Total # 8-Ks	79	90	169

Table 5.4: Number of ‘downs’ and ‘ups’ with unique FEPs in ‘downs’ and ‘ups’, no duplicates (first S&P 500 dataset).

Looking at Table 5.4 more closely, we can see that 74.7% of the ‘downs’ that had one or more recognised FEPs, had one unique FEP and 15.2% had two unique FEPs. With regards the ‘ups’ that had one or more recognised FEPs, 71.1% had one unique FEP and 25.6% had two unique FEPs. Note also how many more ‘ups’ had two unique recognised FEPs, when compared to the ‘downs’ (25.6% vs. 15.2%).

### 5.3.2 Pattern Analysis of FEPs in the Second S&P 500 Dataset

The second S&P 500 dataset relates to the period from 2005-2008. As outlined in Table 4.8, the dataset comprised 2,682,155 million words and 30MB of data. We first determined the total number of FEPs recognised in all files, including duplicate FEPs (i.e. where the same event was mentioned one or more times). Table 5.5 shows the total number of FEPs recognised in all files, the average number of FEPs found per file and the minimum and maximum number of FEPs recognised in any one file. The recogniser identified 258 FEPs in the ‘downs’ and 297 FEPs in the ‘ups’, including duplicates.

Looking at Table 5.5, we note how there were more FEPs recognised in the ‘ups’, when compared to the ‘downs’ (297 vs. 258), which may be partly due to the fact that there were more up 8-Ks. On average, 0.4 FEPs were recognised per file (rounded to one decimal place), although this varied from 0 to 19 FEPs (‘downs’) and 0 to 11 (‘ups’).

	<i>2005-2008 downs (574 8-Ks)</i>	<i>2005-2008 ups (672 8-Ks)</i>
<i>Total number of FEPs recognised in all files, including duplicates</i>	258	297
<i>Average number of FEPs per file</i>	0.4	0.4
<i>Minimum and maximum number of FEPs recognised in any one file</i>	0-19	0-11

Table 5.5: Total, average, and range of FEPs recognised in all ‘downs’ and ‘ups’, including duplicates (second S&P 500 dataset).

We then determined the number of disclosures for which 0, 1, 2, 3... FEPs were identified by the recogniser (see Table 5.6). The relatively high number of instances

of disclosures which had 0 FEP types highlights how difficult it can be to write a definitive financial grammar; for this dataset, our recogniser only succeeded in identifying FEPs in 20.4% of the ‘downs’ and 24.3% of the ‘ups’. For these remaining 8-Ks (117 ‘downs’ and 163 ‘ups’), we also looked at the number of FEPs that were recognised. Table 5.6 also shows the total number of FEPs recognised for the remaining ‘downs’ and ‘ups’, including duplicate FEPs.

# FEPs identified (incl. duplicates)	# Downs (574 8-Ks)	# Ups (672 8-Ks)	Total # disclosures (1246 8-Ks)
0	457 (79.6%)	509 (75.7%)	966 (77.5%)
1	67 (11.7%)	93 (13.8%)	160 (12.8%)
2	26 (4.5%)	40 (6.0%)	66 (5.3%)
3	9 (1.6%)	18 (2.7%)	27 (2.2%)
4	3 (0.5%)	5 (0.7%)	8 (0.6%)
5	3 (0.5%)	3 (0.4%)	6 (0.5%)
6+	9 (1.6%)	4 (0.6%)	13 (1.0%)

Table 5.6: Number of ‘downs’ and ‘ups’ which had 0 or more FEP types (second S&P 500 dataset).

As Table 5.6 shows the number of FEPs recognised, including duplicates, we also had to determine how many *different FEPs types* were identified by the recogniser. Table 5.7 shows that 16 of the 49 FEPs types were identified in the 117 ‘downs’ and there were 134 unique occurrences when duplicate FEPs within individual 8-Ks were removed. In the 163 ‘ups’, 17 of the 49 FEP types were identified and there were 188 unique occurrences when duplicate FEPs within individual 8-Ks were removed. Appendix 24 lists the different FEP types that were recognised for this dataset.

	Downs (117 8-Ks)	Ups (163 8-Ks)
# Different FEP types recognised	16/49 (32.7%)	17/49 (34.0%)
# Unique occurrences (no duplicates within 8-Ks)	134	188

Table 5.7: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (second S&P 500 dataset).

As mentioned in Section 5.3.1, when we discussed the first S&P 500 dataset, we found that 72.8% of the 8-Ks, for which FEPs were recognised, were single-event filings. In the second dataset, when we only considered 8-Ks for which FEPs were recognised (see Table 5.7), and unique occurrences of FEPs (Table 5.8), we found that 87.1% of the 8-Ks were single-event filings (see the second row, fourth column in Table 5.8). However, as mentioned previously, it is possible that this figure could be increased or decreased with a more efficient recogniser; in an ideal situation, we would find at least one FEP type in each 8-K.

Looking at Table 5.8 more closely, we can see that 90% of the ‘downs’ that had one or more recognised FEP, had one unique FEP and 10% had two recognised FEPs. Similarly, 85.3% of the ‘ups’ that had one or more recognised FEP, had one unique FEP and 14.1% had two recognised FEPs.

<i># Unique FEPs identified (no duplicates)</i>	<i>Downs (134 unique FEP occurrences)</i>	<i>Ups (188 unique FEP occurrences)</i>	<i>Downs and Ups (322 unique FEP occurrences)</i>
1	105 (90.0%)	139 (85.3%)	244 (87.1%)
2	7 (10.0%)	23 (14.1%)	30 (10.7%)
3	5 (0.0%)	1 (0.6%)	6 (2.1%)
Total # 8-Ks	117	163	280

Table 5.8: Number of ‘downs’ and ‘ups’ with unique FEPs in ‘downs’ and ‘ups’, no duplicates (second S&P 500 dataset).

### 5.3.3 Discussion of FEP Patterns in Both Datasets

In this section, we examine both datasets together, with a view to identifying possible trends or patterns in the ‘downs’ and ‘ups’. For example, for both datasets, we examine the total number of words, the total number of duplicate FEPs recognised, the different FEP types recognised, and the occurrences of each FEP type, in ‘downs’ and ‘ups’. We also examine the most frequently-occurring FEP types, without duplicates, and the FEP types that occurred together in ‘downs’ and ‘ups’, in both datasets.

With regards the number of words, there were 807,620 words in the first S&P 500 dataset and 2,682,155 words in the second, which represents a significant increase in words (see Tables 4.6 and 4.8 respectively in Chapter 4). We would expect there to be more words in the latter dataset, as there were more than twice as many 8-Ks (1,246 vs. 548). We would also expect there to be more words because companies are now required to file more types of events in their 8-Ks, so it is likely that they will be lengthier. Also of interest is the word count difference between ‘downs’ and ‘ups’, in both datasets (see Table 5.9).

If we compare the number of words in ‘downs’ vs. ‘ups’ in the first S&P 500 dataset (1997-2000 period), we find that there were proportionately more words in the ‘ups’ compared to the downs (455,730 vs. 351,890 words; 292 vs. 256 8-Ks). If we perform the same comparison in the second S&P 500 dataset (2005-2008 period), we find that there were proportionately the *same* number of words in the ‘ups’ and ‘downs’ (1,447,346 vs. 1,234,809 words; 672 vs. 574 8-Ks). This suggests that our ‘ups’ contained more verbose language than the ‘downs’ before the 2004 rule changes but that the ‘ups’ and ‘downs’ were fairly similar (word-count wise) after 2004.

	<i>1997-2000 downs</i>	<i>2005-2008 downs</i>	<i>1997-2000 ups</i>	<i>2005-2008 ups</i>
# of words	351,890	1,234,809	455,730	1,447,346
# of 8-Ks	256	574	292	672

Table 5.9: Number of words in ‘downs’ and ‘ups’ (both datasets).

Also, assuming there is a correlation between the 8-K content and annret (see Chapter 1 for a list of our assumptions), this finding for the first S&P dataset would appear to correlate with the Kohut and Segars (1992) finding that high-performing firms use more verbose language than poor-performing firms (see Section 2.3 for a discussion). On a related note, the same two researchers later found that the readability of reports filed by good performing firms was better than those filed by poor-performing firms; they found that the former tended to use strong, clear and concise language whereas the latter tended to use more jargon and modifiers (Segars and Kohut 2001, also in Section 2.3). This would appear to contradict their earlier finding, as verbose is the opposite of concise, but we must remember that an unreadable document does not

necessarily have to be a lengthy one; it can also be a short one that uses a lot of jargon and modifiers. On a related note, Li (2008) found that the annual reports of poor-performing firms tend to be less readable and that firms with more readable documents tend to have more persistent positive earnings. Finally, Hildebrandt and Snyder (1981) applied the 'Pollyanna Hypothesis' to the writing of annual reports and found that there were significantly more positive words than neutral or negative words, regardless of whether it was a financially good- or bad-year, suggesting that there is a general preference for using positive words in disclosures (see Section 2.3 for a discussion). Once again, assuming there is a correlation between content and content, then this could explain why the 'ups' contained more words before 2004. Finally, it seems clear that the report writing style changed sometime in the intervening period; either the 'ups' became more concise than previously or the 'downs' became more verbose.

We now turn our attention to an examination of the output from the FEP recogniser, for both datasets. Whilst we initially assumed that we would recognise significantly more FEPs in the second S&P 500 dataset, compared to the first S&P 500 dataset, this was not the case. We assumed this because companies were legally required to file more types of events, and more frequently, post 2004 (see Chapter 2 for a discussion). Table 5.10 compares the number of FEPs, including duplicates, recognised in both periods, for 'downs' and 'ups'.

	<i>1997-2000 downs (256 8-Ks)</i>	<i>2005-2008 downs (574 8-Ks)</i>	<i>1997-2000 ups (292 8-Ks)</i>	<i>2005-2008 ups (672 8-Ks)</i>
<i>Total # of FEPs found in all files, including duplicates</i>	274	258	260	297
<i>Average # of FEPs per file</i>	1.1	0.4	0.9	0.4
<i>Min. and max. # of FEPs found in any one file</i>	0-28	0-19	0-18	0-11

Table 5.10: Total, average, and range of FEPs recognised in all 'downs' and 'ups', including duplicates (both datasets).

As we can see in Table 5.10, there were 274 duplicate FEPs recognised in the ‘downs’ in the first dataset, and 258 recognised in the ‘downs’ in the second dataset, even though there were more than twice as many ‘down’ disclosures in the second (574 vs . 256) . However, as we know that the recogniser only succeeded in identifying FEPs in 20.4% of ‘downs’ in the second dataset, compared with 30.9% in the first (see Tables 5.6 and 5.2 respectively), we must take this into consideration. It seems that the recogniser was less able to identify FEPs in the more recent dataset, even though it had a significantly larger collection of 8-Ks to work with.

With regards the ‘ups’, the pattern matches our assumption, in that there were more FEPs recognised in the more recent dataset. However, considering that there were more than twice as many 8-Ks in this period, one might expect there to be significantly more FEPs recognised, but this was not the case. The number of FEPs recognised only increased from 260 to 297. Also, the recogniser only succeeded in finding FEPs in 24.3% of the ‘ups’ in the second dataset, compared with 30.8% in the first, so it is evident that the recogniser encountered more difficulties (compare Tables 5.6 and 5.2 respectively). We must bear in mind, however, that these figures relate to FEPs recognised—it is quite possible that there were more FEPs in these disclosures.

When we compare the average number of FEPs per file, as well as the minimum and maximum number, we also find that the recogniser was less successful in finding FEPs in the second dataset. On average, the recogniser found 1 FEP per file in the first, but 0.4 (none) in the second (see Table 5.10). The maximum number of FEPs recognised also decreased in the second dataset, but this could also be because of more accurate and timely reporting i.e. companies might have started to file single events more frequently than before, rather than file several events in one disclosure.

As already mentioned earlier in this section, we noted that the recogniser found FEPs in a smaller percentage of 8-Ks in the second dataset. One possible suggestion for this is that the language changed significantly in the intervening period, and therefore the recogniser was not able to recognise as many FEPs. It is possible that the marked increase in the number of events that needed to be filed, brought with it a change in

the style of reporting language. Also, as we know that the number of events that had to be filed were increased from 2004, it is quite likely that some of the new events were not catered for in the FEP recogniser. This limitation will be discussed further in Chapter 7.

With regards to the number of different FEP types recognised, there was not much difference between the two periods (see Table 5.11). Perhaps more interesting is the fact that the number of unique occurrences, when duplicate FEPs were removed, did increase in the second dataset, which answers one of our secondary research questions (see Chapter 1). Whilst in Table 5.10 we found that the number of *duplicate* FEPs decreased in the second dataset, here we find that the number of *unique* FEPs increased (110 increased to 134 and 119 increased to 188). One possible reason for this could be that there was less repetition of each event, as auditors became more conscientious about their reporting style after the 2004 changes.

	<i>1997-2000 downs (256 8-Ks)</i>	<i>2005-2008 downs (574 8-Ks)</i>	<i>1997-2000 ups (292 8-Ks)</i>	<i>2005-2008 ups (672 8-Ks)</i>
<i># Different FEP types recognised</i>	18/49	16/49	19/49	17/49
<i># Unique occurrences (no duplicates) within 8-Ks</i>	110	134	119	188

Table 5.11: Number of different FEP types recognised and number of unique occurrences in ‘downs’ and ‘ups’ (both datasets).

Another area worth examining is the number of single-event filings. As mentioned in Sections 5.3.2 and 5.3.3, 72.8% and 87.1% of the 8-Ks that had FEPs, were single-event filings (first and second dataset respectively). This correlates with another one of our research questions (see Chapter 1), which is that we expect the more recent 8-Ks to be more focused and less ‘noisy’, as a result of the changed regulations. However, we must remember that these figures might be increased or decreased with a more efficient recogniser—it is quite possible that there were more FEP types in these 8-Ks, in both datasets.

When we examine the different FEP types found by the recogniser, we can see that the majority of FEP types that appeared in the first dataset also appeared in the second. Table 5.12 lists the FEP types recognised in both datasets, where Y stands for ‘Yes, it was recognised’ and N stands for ‘No, not recognised’. FEP types highlighted in bold type are discussed here in the text. Only `fep_accountant_dismissal`, `fep_dividend_distribution`, and `fep_private_placement`, which were recognised in the second dataset, were not recognised in the initial dataset. In the second dataset, only `fep_stock_offering`, `fep_stock_option_agreement_or_plan`, and `fep_will_or_has_generated_loss` were not recognised (they were recognised in the first). The only FEP type that appears to be correlated with an increase in share price is `fep_accountant_appointment`, which only appeared in the ‘ups’ (both time periods). We might expect a new appointment to be correlated with good news, especially if there has been an issue with a previous accountant.

<i>FEP type</i>	<i>1997-2000 downs</i>	<i>2005-2008 downs</i>	<i>1997-2000 ups</i>	<i>2005-2008 ups</i>
<b>fep_accountant_appointment</b>	N	N	<b>Y</b>	<b>Y</b>
<b>fep_accountant_dismissal</b>	<b>N</b>	Y	N	Y
<code>fep_acq_ag_and_plan_of_merger_or_reorg</code>	Y	Y	Y	Y
<code>fep_allocation_writeoff_or_amortization_of_goodwill</code>	Y	Y	Y	N
<code>fep_amend_rights_plan_issue_offer_or_agreement</code>	N	Y	Y	N
<code>fep_could_would_can_expect_material_adverse_effect</code>	Y	Y	Y	N
<b>fep_dividend_distribution</b>	<b>N</b>	N	N	Y
<code>fep_indenture</code>	Y	N	Y	Y
<code>fep_new_personnel_or_promotions</code>	Y	Y	Y	Y
<b>fep_private_placement</b>	<b>N</b>	N	<b>N</b>	Y
<code>fep_purchase_agreement</code>	Y	Y	Y	Y
<code>fep_remain_as_personnel</code>	N	Y	Y	Y
<code>fep_resignation_or_leaving</code>	Y	Y	Y	Y
<code>fep_rights_plan_issue_offer_or_agreement</code>	Y	Y	Y	N
<code>fep_securities_purchase_agreement</code>	Y	Y	Y	Y
<b>fep_stock_offering</b>	Y	<b>N</b>	N	<b>N</b>

<b>fep_stock_option_agreement_or_plan</b>	Y	N	N	N
fep_stock_or_note_conversion	Y	Y	Y	Y
fep_stock_purchase_agreement	Y	Y	Y	Y
fep_stock_split	Y	Y	Y	N
fep_to_amend_general_ag_report_doc_or_info	Y	Y	Y	Y
fep_will_or_has_generated_income	Y	Y	Y	Y
<b>fep_will_or_has_generated_loss</b>	Y	N	Y	N
fep_will_or_has_generated_revenue	Y	N	Y	Y

Table 5.12: Occurrences of each FEPT type in ‘downs’ and ‘ups’ (both datasets).

Looking more closely at the five events that had the highest number of occurrences in both time periods (see Table 5.13), we found that the `fep_acq_ag_and_plan_of_merger_or_reorg` occurred in a fairly even number of ‘ups’ and ‘downs’, in both datasets. Merger and acquisition agreements can have different implications for companies, depending on whether they are the acquiring company or the company being acquired. Also, specific details regarding a merger can have different implications for shareholders as they often have associated stock changes, as we will see shortly.

The `fep_new_personnel_or_promotions` was recognised in more ‘ups’ than ‘downs’ and there were more occurrences in the second (more recent) dataset. We would expect this to be the case, as this event is more likely good news than bad news and there were more 8-Ks in the second dataset.

The `fep_purchase_agreement` was recognised in more ‘ups’ than ‘downs’, but there were less occurrences recognised in the second dataset, which is surprising seeing as there were more 8-Ks. The `fep_stock_or_note_conversion` was recognised in a fairly equal number of ‘ups’ and ‘downs’ but in slightly less 8-Ks in the second dataset. Finally, the `fep_will_or_has_generated_income` was recognised in more ‘ups’ than ‘downs’, which was expected. As there were more occurrences in the second dataset, this correlates with our assumption that there would be more occurrences in a larger dataset.

<i>FEP type</i>	<i>1997-2000 downs</i>	<i>2005-2008 downs</i>	<i>1997-2000 ups</i>	<i>2005-2008 ups</i>
fep_acq_ag_and_plan_of_merger_or_reorg	21	19	22	18
fep_new_personnel_or_promotions	34	59	26	90
fep_purchase_agreement	5	4	12	4
fep_stock_or_note_conversion	9	8	11	8
fep_will_or_has_generated_income	13	17	11	22

Table 5.13: Most frequently-occurring FEP types in ‘downs’ and ‘ups’, no duplicates (both datasets).

Looking at the FEP combinations which occurred together in more than one 8-K (Table 5.14), some combinations were only recognised in the first dataset; for example, the `fep_allocation_writeoff_or_amortization_of_goodwill` and `fep_new_personnel_or_promotions` as well as `fep_new_personnel_or_promotions` and `fep_stock_or_note_conversion`. Both combinations appeared in an equal number of ‘ups’ and ‘downs’, in both datasets. The combination of `fep_acq_ag_and_plan_of_merger_or_reorg` and `fep_new_personnel_or_promotions`, appeared in more ‘downs’ overall but also appeared more in the first dataset. No mergers and new personnel were recognised in the ‘ups’ in the second dataset, which could be because of inefficiencies in the recogniser, or perhaps more of these events were considered bad news than good news. Other combinations appear to be correlated with decreases in share prices, as they only appeared in ‘downs’: the `fep_acq_ag_and_plan_of_merger_or_reorg` and `fep_to_amend_general_ag_report_doc_or_info`.

The `fep_acq_ag_and_plan_of_merger_or_reorg` and `fep_stock_or_note_conversion` combination appeared more in the first dataset but appeared in a fairly equal number of ‘downs’ and ‘ups’, in each dataset. The combination of `fep_new_personnel_or_promotions` and `fep_remain_as_personnel` appeared in more ‘ups’ than ‘downs’, and we would probably expect this to be more good news than bad news.

<i>FEP type</i>	<i>1997-2000 downs</i>	<i>2005-2008 downs</i>	<i>1997-2000 ups</i>	<i>2005-2008 ups</i>
fep_acq_ag_and_plan_of_merger_or_reorg fep_new_personnel_or_promotions	6	3	5	0
fep_acq_ag_and_plan_of_merger_or_reorg fep_to_amend_general_ag_report_doc_or_info	4	2	0	0
fep_acq_ag_and_plan_of_merger_or_reorg fep_stock_or_note_conversion	6	3	8	2
fep_new_personnel_or_promotions fep_remain_as_personnel	0	2	2	4
fep_allocation_writeoff_or_amortization_of_goodwill fep_new_personnel_or_promotions	3	0	3	0
fep_new_personnel_or_promotions fep_stock_or_note_conversion	2	0	2	0

Table 5.14: FEP types that occurred together in ‘downs’ and ‘ups’ (both datasets).

Our final analysis of both datasets relates to duplicate FEPs. Was there more repetition of any given FEP in the first or second dataset? Table 5.15 combines the data from Tables 5.8 and 5.16 and shows the most frequently-occurring FEP types for both datasets. The value ‘N/A’ is used where that FEP type was not one of the most frequently-occurring FEP types for a particular dataset but this does not necessarily mean the FEP type did not occur—it just might not have occurred as frequently as others. The *fep\_new\_personnel\_or\_promotions* was repeated in slightly more ‘downs’ but it was repeated more than twice as often in the second (more recent) dataset. Whilst we would expect to find more occurrences of the FEP in the second dataset, as there are more 8-Ks, we would expect it to appear more in ‘ups’ than ‘downs’. It is possible that this FEP type was repeated more in ‘downs’, to partly disguise other negative news or at least reduce the likely impact. Similar tactics—which Malkiel (2007) refers to as “creative accounting procedures” (p.155)—have been highlighted in previous studies (e.g. see Kohut and Segars 1992 and Li 2008 in Chapter 2).

The *fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg* appeared in more ‘downs’ but it also appeared more in the first dataset, which could be because the report writers used more concise writing strategies in the more recent dataset (i.e. perhaps

they did not repeat the event as often). On the other hand, it could be due to inefficiencies in the recogniser, as less FEPs were recognised overall in the second dataset. The *fep\_will\_or\_has\_generated\_income* occurred more frequently in the second dataset and in more ‘ups’ than ‘downs’, which we would expect. The *fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill* and *fep\_securities\_purchase\_agreement* were repeated more frequently in the first dataset, but this could possibly be due to changes in writing style, rather than a reduction in the occurrences of that FEP type in the second dataset. For example, in another study, Loughran et al (2008) found that ethics-related terms appeared in more Form 10-Ks once companies were legally required to disclose ethics-related events, which suggests that the SEC changes forced report writers to modify their writing style. Kohut and Segars (1992) found that the direct style was used by companies disclosing good news and the indirect style was used by companies disclosing bad news.

<i>FEP type</i>	<i>1997-2000 downs</i>	<i>2005-2008 downs</i>	<i>1997-2000 ups</i>	<i>2005-2008 ups</i>
<i>fep_new_personnel_or_promotions</i>	74	141	43	163
<i>fep_acq_ag_and_plan_of_merger_or_reorg</i>	100	40	91	29
<i>fep_will_or_has_generated_income</i>	N/A	21	N/A	30
<i>fep_allocation_writeoff_or_amortization_of_goodwill</i>	20	N/A	27	N/A
<i>fep_securities_purchase_agreement</i>	20	N/A	3	N/A

Table 5.15: Most frequently-occurring FEP types in ‘downs’ and ‘ups’ (both datasets).

## 5.4 Summary

It is clear from the pattern analyses in this chapter that our FEP recogniser needs more work. Firstly, the phrases incorporated into the recogniser need to be developed further, so the recogniser can identify one or more FEPs in a greater number of disclosures. This may necessitate reducing precision in favour of a higher recall; for example, if we try to recognise shorter FEP phrases, the rate of recall may increase but the precision will likely decrease. Secondly, some of the existing phrases need to be refined, so the recogniser does not make mistakes. We believe that stock-related events, in particular, can be very difficult to classify and may warrant further subdivision.

If disclosures filed after 2004 typically only refer to one type of FEP and there are many different types of FEP (this seems to be the case, judging by our manual analysis of 28 disclosures), then our classifier would need a very large dataset to learn patterns from this data. At the end of Section 4.4, we mentioned that we made a list of 1,568 frequently-occurring keywords (excluding stop words and other non-interesting words), whilst we were preparing the FEP lists. After reviewing the pattern analyses findings, we decided to use these keywords as *additional* features in the classification process to give the classifiers the best possible chance. In the next chapter, we will present the results from our classification experiments with C4.5 (Section 6.2) and SVM-Light (Section 6.3), where we used FEPs and keywords as document content features. We will also present the results from a number of experiments which were used as baseline approaches to compare with these experiments (see Sections 6.4 to 6.6).

# Chapter 6: Automatic Classification of 8-K Disclosures using Various Document Content Features

## 6.1 Outline

In Chapter 3, we briefly introduced the topics of automatic text classification and categorisation. We then focused extensively on various methods that have been used for the automatic analysis of financial documents, including rule induction, neural networks, support vector machines, and Bayesian methods. Whilst these were the most frequently-used methods, statistical methods, language modelling, multiple discriminant analysis, k-nearest neighbour, and genetic algorithms were also used.

In Chapter 4, we presented our rationale for focusing on one type of financial document—the Form 8-K—and we described the features of our datasets. In Chapter 5, we described the features of our prototype financial event phrase (FEP) recogniser and highlighted some FEP patterns that emerged in the different datasets of 8-Ks.

In this chapter, we present the results from our experiments using two different methods (decision trees and support vector machines) for the classification of Form 8-Ks by likely share price response, using FEPs and keywords as content features. For each method, we first describe the systems we chose (C4.5 and SVM-Light), before presenting the detailed results of our experiments. We then analyse the results for each experiment. We also present and analyse the results from two alternative approaches that were used as baseline approaches: classification using automatically-extracted n-gram features (where  $n=5$ ) (see Sections 6.4 and 6.5) and classification using a bag-of-words (Section 6.6). Finally, we provide a summary of the chapter.

## 6.2 Decision Tree Classification using Financial Event Phrases and Keywords

In this section, we first present an overview of C4.5, the decision tree classification method we adopted. We then present the results from a number of C4.5 experiments in Sections 6.2.2 to 6.2.5:

- Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5).
- Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5).
- Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5).
- Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5).

In Section 6.2.6, we will analyse and discuss the results from all four experiments.

### 6.2.1 Background to Decision Trees and C4.5

C4.5 is one of a family of well-known decision tree learning algorithms that also includes its predecessor, ID3 (Mitchell 1997). We decided to use a decision tree learning algorithm because these kinds of algorithms have been used quite successfully in the financial domain (see, for example, Koppel and Shtrimberg 2004, Peramunetilleke and Wong 2002, Wüthrich et al 1998, and Koh and Low 2004 in Chapter 3). In a recent study by Robles-Granda and Belik (2010), C4.5 was found to perform better during the training phase when predicting stock volatility, than an artificial neural network with back propagation and a Naïve Bayes classifier. Whilst C4.5 proved to be less stable during the classification phase (without pruning), it still achieved a better level of accuracy overall. Other reasons for choosing C4.5 include its processing speed, its ability to handle noisy textual data (Harries and Horn 1995), and our familiarity with it.

C4.5 is a suite of programs that construct classification models by examining numerous recorded classifications<sup>39</sup> (Quinlan 1993). These models aim to discover

---

<sup>39</sup> C5.0 has replaced C4.5 since we initiated these experiments.

and analyse patterns found in a number of cases. Each case must be represented by a series of non-category features or attributes, with each attribute having either a discrete or numeric value. Each case is then assigned a particular category or class, which must be predefined—this is known as supervised learning—and each case must belong to only one class.

In addition to ensuring that there are more cases than classes, it is also necessary to ensure there is sufficient data. The amount of data required is affected by numerous factors, including the number of attributes, the number of classes, and the complexity of the classification problem (Quinlan 1993). Lewis has done extensive work on feature selection and extraction (see, for example, Lewis 1992a). Robles-Granda and Belik (2010) found that reducing the number of features did not significantly impact the classification accuracy and that the noisy nature of their data did not affect the stability of C4.5.

Cases are sometimes split into training and test data, with the former used to construct the decision tree and/or rule classification models and the latter used to test the accuracy of the models (Quinlan 1993; Berthold and Hand 2003). However, this is not always ideal, especially where relatively small datasets are used (e.g. less than a few hundred). An alternative approach is to use K-fold cross-validation, whereby the data is split into K folds (Quinlan 1993). Folds are usually evenly sized, although there can be minor differences, depending on the number of folds and the size of the dataset. For each of the K folds, the remaining K-1 folds are used as training data. K is withheld as test data, thereby ensuring that each fold is used only once as test data (Quinlan 1993; Berthold and Hand 2003). Provided the number of folds is not too small—ten-fold is usually recommended—the average error rate over the unseen test cases should be a good predictor of the error rate of a model built from all the data (Quinlan 1993).

Using our financial classification problem as an example, each Form 8-K disclosure is a case. Each case either contains, or does not contain, one or more attributes or features (financial event phrases, named entities, types of financial object, and keywords). The class (up or down) is specified for both the training and test data but

C4.5 uses the training models to evaluate the accuracy on the previously-unseen test data, the latter being presented at each fold. C4.5 returns an error rate for the unseen test data, which can be compared with the error rate for the training data. Usually, the model performs better on the training data but the goal is to achieve as low an error as possible on the unseen test data.

When preparing data for C4.5, three files need to be prepared (Quinlan 1993). The first—the .names file—lists each possible class (up or down, in our study). It also lists the values for each attribute (see Figure 6.1 for some sample attributes and their values). The second file—the .data training file—comprises the values of each attribute for each training case, and the third file—the .test file—comprises the same type of data for each test case (see Figures 6.2 and 6.3 respectively). As this is supervised learning, we also had to provide the actual classification (up or down) so the algorithm could determine its performance accuracy on both the training and test data.

```
UP,DOWN.
```

```
ability: TRUE,FALSE.
```

```
absence: TRUE,FALSE.
```

```
...
```

```
written: TRUE,FALSE.
```

```
year: TRUE,FALSE.
```

```
years: TRUE,FALSE.
```

```
fep_accountant_dismissal: TRUE,FALSE.
```

```
fep_accountant_appointment: TRUE,FALSE.
```

```
...
```

```
fep_owes_compensation: TRUE,FALSE.
```

```
fep_to_amend_general_ag_report_doc_or_info: TRUE,FALSE.
```

```
type_accountant_or_accountant_name: TRUE,FALSE.
```

```
type_appointment_or_promotion: TRUE,FALSE.
```

```
...
```

Figure 6.1: Sample content for the C4.5 .names file (keywords, FEPs, named entities, and types of financial object).

```
TRUE,FALSE,FALSE,...,TRUE,DOWN.
```

```
FALSE,TRUE,TRUE,...,FALSE,UP.
```

```
...
```

Figure 6.2: Sample content for the C4.5 .data training file (two cases).

```
TRUE,TRUE,FALSE,...,FALSE,DOWN.
```

```
TRUE,FALSE,TRUE,...,FALSE,UP.
```

```
...
```

Figure 6.3: Sample content for the C4.5 .test test file (two cases).

C4.5 classification output can take the form of decision trees or production rules, which describe the class by referring to the values of particular attributes (Quinlan 1993). Separate trees or rules are generated for each fold, in the case of K-fold cross validation.

C4.5 decision trees typically take the following format (Quinlan 1993): A tree comprises a number of nodes, with each node corresponding to a non-category attribute. Each node has one or more branches, each leading to a possible value of that attribute. At the end each branch, there is a leaf, specifying the value of the category attribute (or class). Each node should possess the non-category attribute with the *most* information at that point; however, only attributes not yet considered at that point (from the root), should be included at a node. The process is best described by Quinlan (1993) as follows:

*“A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered. At each nonleaf decision node, the case’s outcome for the test at the node is determined and attention shifts to the root of the subtree corresponding to this outcome. When this process finally (and inevitably) leads to a leaf, the class of the case is predicted to be that recorded at the leaf” (pp.5-6).*

A number presented beside each leaf specifies the number of training instances, out of the total number of cases evaluated, which belong to that leaf or path in the tree. In

some cases, this number is followed by a second number, which specifies how many of those classifications were in fact errors (Quinlan 1993). The usually lengthy decision trees are also pruned or simplified, by replacing subtrees with leaf nodes; however, pruning only happens if “the expected error rate in the subtree is greater than in the single leaf” (Ingargiola date unknown). Both the original and pruned trees are presented in the output.

In addition to presenting the trees, C4.5 also provides summary details outlining the evaluation on training data: it presents the size of the trees (number of nodes), the number of errors using both the original and pruned trees, as well as the estimated error percentage for the unseen test data. C4.5 then presents the equivalent summary details for the test data, so it is possible to compare the estimated error percentage with the actual error percentage. Finally, it presents a confusion matrix (Quinlan 1993) based on the pruned tree, which details the number of correct and incorrect classifications for each class.

Production rules can also be derived from C4.5 decision trees. These are written from the root to the leaf, usually in the form *if... then...* rules and they are usually easier to interpret than the lengthy unpruned decision trees from which they are derived. One class is also designated as a default (Quinlan 1993). In our research, we briefly evaluated the performance of the rules, but our main focus was on the performance of the decision trees and the confusion matrices that accompany them. In Sections 6.2.2 to 6.2.5, we will present the results from our C4.5 experiments and in Section 6.2.6 we will analyse and discuss the results.

### **6.2.2 Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5)**

This experiment used 169 disclosures from the 1997-2000 dataset and comprised 90 ‘ups’ and 79 ‘downs’. Each of these disclosures had one or more recognised FEPs within them, derived from our prototype financial event phrase recogniser. We used 1,635 features for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects. As discussed in Section 5.5, we decided to introduce keywords into the classification process due to the poor recognition performance of our FEP recogniser.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 2.9% error rate on the training data and an average 47.9% error rate (or 52.1% accuracy rate) on the unseen test data (see the last two rows in Table 6.1). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	61	4 (2.6%)	51	5 (3.3%)	(23.9%)
	<b>61</b>	<b>6 (35.3%)</b>	<b>51</b>	<b>6 (35.3%)</b>	<b>(23.9%)</b>
2	57	6 (3.9%)	57	6 (3.9%)	(25.8%)
	<b>57</b>	<b>6 (35.3%)</b>	<b>57</b>	<b>6 (35.3%)</b>	<b>(25.8%)</b>
3	59	2 (1.3%)	59	2 (1.3%)	(23.9%)
	<b>59</b>	<b>9 (52.9%)</b>	<b>59</b>	<b>9 (52.9%)</b>	<b>(23.9%)</b>
4	57	3 (2.0%)	57	3 (2.0%)	(23.8%)
	<b>57</b>	<b>8 (47.1%)</b>	<b>57</b>	<b>8 (47.1%)</b>	<b>(23.8%)</b>
5	61	4 (2.6%)	61	4 (2.6%)	(25.5%)
	<b>61</b>	<b>10 (58.8%)</b>	<b>61</b>	<b>10 (58.8%)</b>	<b>(25.5%)</b>
6	59	1 (0.7%)	59	1 (0.7%)	(23.0%)
	<b>59</b>	<b>8 (47.1%)</b>	<b>59</b>	<b>8 (47.1%)</b>	<b>(23.0%)</b>
7	51	9 (5.9%)	45	10 (6.6%)	(24.5%)
	<b>51</b>	<b>9 (52.9%)</b>	<b>45</b>	<b>10 (58.8%)</b>	<b>(24.5%)</b>
8	65	0 (0.0%)	63	0 (0.0%)	(23.6%)
	<b>65</b>	<b>8 (47.1%)</b>	<b>63</b>	<b>8 (47.1%)</b>	<b>(23.6%)</b>
9	57	6 (3.9%)	55	6 (3.9%)	(25.0%)
	<b>57</b>	<b>9 (52.9%)</b>	<b>55</b>	<b>9 (52.9%)</b>	<b>(25.0%)</b>
10	65	4 (2.6%)	51	7 (4.6%)	(24.9%)
	<b>65</b>	<b>7 (43.8%)</b>	<b>51</b>	<b>7 (43.8%)</b>	<b>(24.9%)</b>
<i>Train (average):</i>	59.2	3.9 (2.6%)	55.8	4.4 ( <b>2.9%</b> )	(24.4%)
<i>Test (average):</i>	59.2	8.0 (47.3%)	55.8	8.1 ( <b>47.9%</b> )	(24.4%)

Table 6.1: Number of errors and tree sizes for the 1997-2000 dataset: subset 1 (C4.5).

We also briefly examined the performance of the production rules derived from the original decision trees (Quinlan 1993) and found that, on average, there was a 21.0% error rate on the training data and an average 47.8% error rate (or 52.2% accuracy) on the unseen data. If we wish to compare the average performance of the rules with the average performance of the trees, we must compare the former with the performance of the unpruned or original tree. Whilst the training error with the rules (21.0%) was significantly higher than the error with the original tree (2.6%, see the second last row and third column in Table 6.1), the error rate on the unseen data was just marginally higher (47.8% vs. 47.3% respectively). In this experiment, the rules did not yield improved performance over the original decision trees, on the unseen data. As outlined earlier, the rest of this section will focus on the output from the trees.

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.2 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 9 test cases as UP (second row, last column) and 8 of these were correctly classified (second row, second column). C4.5 classified 8 cases as DOWN (second row, last column) but only 3 of these were DOWN (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>8</b>	1	(a): class UP	9
	5	<b>3</b>	(b): class DOWN	8
<i>2</i>	<b>6</b>	3	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>3</i>	<b>4</b>	5	(a): class UP	9
	4	<b>4</b>	(b): class DOWN	8
<i>4</i>	<b>4</b>	5	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>5</i>	<b>5</b>	4	(a): class UP	9
	6	<b>2</b>	(b): class DOWN	8
<i>6</i>	<b>6</b>	3	(a): class UP	9
	5	<b>3</b>	(b): class DOWN	8
<i>7</i>	<b>2</b>	7	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8

8	<b>3</b>	6	(a): class UP	9
	2	<b>6</b>	(b): class DOWN	8
9	<b>4</b>	5	(a): class UP	9
	4	<b>4</b>	(b): class DOWN	8
10	<b>6</b>	3	(a): class UP	9
	4	<b>3</b>	(b): class DOWN	7
<i>Total</i>	<b>48</b>	42	(a): class UP	90
	39	<b>40</b>	(b): class DOWN	79

Table 6.2: Confusion matrix for the 1997-2000 dataset: subset 1 (C4.5).

If we add all the correct UP classifications, we find that 48/90 UP classifications were correctly classified (53.3%). If we compute a similar value for the DOWNS, we find that 40/79 (50.6%) were correctly classified. A discussion of these results can be found in Section 6.2.6.

### 6.2.3 Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5)

This experiment used 280 disclosures from the 2005-2008 dataset, comprising 163 ‘ups’ and 117 ‘downs’. Each of these disclosures had one or more recognised FEPs within them, derived from our prototype FEP recogniser. Like the previous experiment, we used 1,635 attributes for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 7.7% error rate on the training data and an average 47.5% error rate (or 52.5% accuracy rate) on the unseen test data (see the last two rows in Table 6.3). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

We also briefly examined the performance of the production rules derived from the original decision trees and found that, on average, there was a 24.4% error rate on the training data and an average 47.1% error rate (or 52.9% accuracy) on the unseen test data. If we wish to compare the average performance of the rules with the average performance of the trees, we must compare the former with the performance of the unpruned or original tree. Whilst the training error with the rules (24.4%) was

significantly higher than the training error with the original tree (5.1%, see the second last row and third column in Table 6.3), the error rate on the unseen test data was lower (47.1% vs. 49.3% respectively). In this experiment, the rules performed slightly better on the unseen test data but for consistency across experiments, the remainder of this section will focus on the output from the decision trees.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	107	20 (7.9%)	59	33 (13.1%)	(27.9%)
	<b>107</b>	<b>9 (32.1%)</b>	<b>59</b>	<b>9 (32.1%)</b>	<b>(27.9%)</b>
2	99	11 (4.4%)	93	12 (4.8%)	(26.2%)
	<b>99</b>	<b>17 (60.7%)</b>	<b>93</b>	<b>16 (57.1%)</b>	<b>(26.2%)</b>
3	105	19 (7.5%)	63	32 (12.7%)	(28.3%)
	<b>105</b>	<b>15 (53.6%)</b>	<b>63</b>	<b>12 (42.9%)</b>	<b>(28.3%)</b>
4	109	4 (1.6%)	93	4 (1.6%)	(22.8%)
	<b>109</b>	<b>17 (60.7%)</b>	<b>93</b>	<b>16 (57.1%)</b>	<b>(22.8%)</b>
5	103	17 (6.7%)	63	27 (10.7%)	(26.5%)
	<b>103</b>	<b>10 (35.7%)</b>	<b>63</b>	<b>11 (39.3%)</b>	<b>(26.5%)</b>
6	105	14 (5.6%)	81	15 (6.0%)	(24.9%)
	<b>105</b>	<b>13 (46.4%)</b>	<b>81</b>	<b>13 (46.4%)</b>	<b>(24.9%)</b>
7	99	9 (3.6%)	87	9 (3.6%)	(24.0%)
	<b>99</b>	<b>13 (46.4%)</b>	<b>87</b>	<b>13 (46.4%)</b>	<b>(24.0%)</b>
8	111	8 (3.2%)	103	8 (3.2%)	(27.0%)
	<b>111</b>	<b>17 (60.7%)</b>	<b>103</b>	<b>16 (57.1%)</b>	<b>(27.0%)</b>
9	95	13 (5.2%)	73	23 (9.1%)	(26.7%)
	<b>95</b>	<b>14 (50.0%)</b>	<b>73</b>	<b>15 (53.6%)</b>	<b>(26.7%)</b>
10	107	14 (5.6%)	63	30 (11.9%)	(27.3%)
	<b>107</b>	<b>13 (46.4%)</b>	<b>63</b>	<b>12 (42.9%)</b>	<b>(27.3%)</b>
<i>Train (average):</i>	104.0	12.9 (5.1%)	77.8	19.3 (7.7%)	(26.2%)
<i>Test (average):</i>	104.0	13.8 (49.3%)	77.8	13.3(47.5%)	(26.2%)

Table 6.3: Number of errors and tree sizes for the 2005-2008 dataset: subset 2 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.4 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 16 test cases as UP (second row, last column) and 15 of these were correctly classified (second row, second column). C4.5 classified 12 cases as DOWN (second row, last column) but only 4 were correctly classified (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>15</b>	1	(a): class UP	16
	8	<b>4</b>	(b): class DOWN	12
<i>2</i>	<b>9</b>	7	(a): class UP	16
	9	<b>3</b>	(b): class DOWN	12
<i>3</i>	<b>10</b>	6	(a): class UP	16
	6	<b>6</b>	(b): class DOWN	12
<i>4</i>	<b>9</b>	7	(a): class UP	16
	9	<b>3</b>	(b): class DOWN	12
<i>5</i>	<b>10</b>	6	(a): class UP	16
	5	<b>7</b>	(b): class DOWN	12
<i>6</i>	<b>8</b>	8	(a): class UP	16
	5	<b>7</b>	(b): class DOWN	12
<i>7</i>	<b>10</b>	6	(a): class UP	16
	7	<b>5</b>	(b): class DOWN	12
<i>8</i>	<b>9</b>	8	(a): class UP	17
	8	<b>3</b>	(b): class DOWN	11
<i>9</i>	<b>11</b>	6	(a): class UP	17
	9	<b>2</b>	(b): class DOWN	11
<i>10</i>	<b>12</b>	5	(a): class UP	17
	7	<b>4</b>	(b): class DOWN	11
<i>Total</i>	<b>103</b>	60	(a): class UP	163
	73	<b>44</b>	(b): class DOWN	117

Table 6.4: Confusion matrix for the 2005-2008 dataset: subset 2 (C4.5).

If we add all the correct UP classifications, we find that 103/163 UP classifications were correctly classified (63.2%). If we compute a similar value for the DOWNS, we find that 44/117 (37.6%) were correctly classified. A discussion of these results can be found in Section 6.2.6.

#### 6.2.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5)

This experiment used all the data from the previous two experiments (Sections 6.2.2. and 6.2.3). Merging this data provided us with a larger number of 8-Ks with one or more recognised FEPs. There were 449 disclosures, comprising 253 ‘ups’ and 196 ‘downs’. Like the previous two experiments, we used 1,635 attributes for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 1.6% error rate on the training data and an average 47.2% error rate (or 52.8% accuracy rate) on the unseen test data (see the last two rows in Table 6.5). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

We also briefly examined the performance of the production rules derived from the original decision trees and found that, on average, there was a 35.8% error rate on the training data and an average 44.1% error rate (or 55.9% accuracy) on the unseen test data. If we wish to compare the average performance of the rules with the average performance of the trees, we must compare the former with the performance of the unpruned or original tree. Whilst the training error with the rules (35.8%) was significantly higher than the training error with the original tree (0.9%, see the second last row and third column in Table 6.5), the error rate on the unseen test data was lower (44.1% vs. 48.6% respectively). In this experiment, the rules performed slightly better on the unseen test data but for the reasons outlined earlier, the remainder of this section focuses on the output from the decision trees.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
<i>1</i>	199	3 (0.7%)	175	8 (2.0%)	(26.8%)
	<b>199</b>	<b>16 (35.6%)</b>	<b>175</b>	<b>16 (35.6%)</b>	<b>(26.8%)</b>
<i>2</i>	195	5 (1.2%)	179	7 (1.7%)	(27.0%)

	<b>195</b>	<b>25 (55.6%)</b>	<b>179</b>	<b>23 (51.1%)</b>	<b>(27.0%)</b>
3	201	2 (0.5%)	177	6 (1.5%)	(26.5%)
	<b>201</b>	<b>17 (37.8%)</b>	<b>177</b>	<b>17 (37.8%)</b>	<b>(26.5%)</b>
4	193	4 (1.0%)	175	7 (1.7%)	(26.6%)
	<b>193</b>	<b>26 (57.8%)</b>	<b>175</b>	<b>26 (57.8%)</b>	<b>(26.6%)</b>
5	195	4 (1.0%)	171	8 (2.0%)	(26.5%)
	<b>195</b>	<b>27 (60.0%)</b>	<b>171</b>	<b>25 (55.6%)</b>	<b>(26.5%)</b>
6	189	4 (1.0%)	175	5 (1.2%)	(25.8%)
	<b>189</b>	<b>18 (40.0%)</b>	<b>175</b>	<b>17 (37.8%)</b>	<b>(25.8%)</b>
7	205	3 (0.7%)	179	8 (2.0%)	(27.4%)
	<b>205</b>	<b>19 (42.2%)</b>	<b>179</b>	<b>20 (44.4%)</b>	<b>(27.4%)</b>
8	209	3 (0.7%)	175	5 (1.2%)	(26.3%)
	<b>209</b>	<b>26 (57.8%)</b>	<b>175</b>	<b>24 (53.3%)</b>	<b>(26.3%)</b>
9	171	5 (1.2%)	157	7 (1.7%)	(24.2%)
	<b>171</b>	<b>19 (42.2%)</b>	<b>157</b>	<b>19 (42.2%)</b>	<b>(24.2%)</b>
10	181	5 (1.2%)	169	6 (1.5%)	(25.4%)
	<b>181</b>	<b>25 (56.8%)</b>	<b>169</b>	<b>25 (56.8%)</b>	<b>(25.4%)</b>
<i>Train</i> (average):	193.8	3.8 (0.9%)	173.2	6.7 ( <b>1.6%</b> )	(26.2%)
<i>Test</i> (average):	193.8	21.8 (48.6%)	173.2	21.2 ( <b>47.2%</b> )	(26.2%)

Table 6.5: Number of errors and tree sizes for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.6 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 25 test cases as UP (second row, last column) and 17 of these were correctly classified (second row, second column). C4.5 classified 20 cases as DOWN (second row, last column) and 12 were correctly classified (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>17</b>	8	(a): class UP	25
	8	<b>12</b>	(b): class DOWN	20
<i>2</i>	<b>14</b>	11	(a): class UP	25
	12	<b>8</b>	(b): class DOWN	20
<i>3</i>	<b>17</b>	8	(a): class UP	25
	9	<b>11</b>	(b): class DOWN	20
<i>4</i>	<b>10</b>	15	(a): class UP	25
	11	<b>9</b>	(b): class DOWN	20
<i>5</i>	<b>12</b>	13	(a): class UP	25
	12	<b>8</b>	(b): class DOWN	20
<i>6</i>	<b>15</b>	10	(a): class UP	25
	7	<b>13</b>	(b): class DOWN	20
<i>7</i>	<b>17</b>	8	(a): class UP	25
	12	<b>8</b>	(b): class DOWN	20
<i>8</i>	<b>13</b>	13	(a): class UP	26
	11	<b>8</b>	(b): class DOWN	19
<i>9</i>	<b>16</b>	10	(a): class UP	26
	9	<b>10</b>	(b): class DOWN	19
<i>10</i>	<b>13</b>	13	(a): class UP	26
	12	<b>6</b>	(b): class DOWN	18
<i>Total</i>	<b>144</b>	109	(a): class UP	253
	103	<b>93</b>	(b): class DOWN	196

Table 6.6: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).

If we add all the correct UP classifications, we find that 144/253 UP classifications were correctly classified (56.9%). If we compute a similar value for the DOWNS, we find that 93/196 (47.4%) were correctly classified. A discussion of these results can be found in Section 6.2.6.

### 6.2.5 Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5)

This experiment used 1,246 disclosures that were filed for the 50 S&P500 companies, during the period 2005-2008. Only the disclosures (cases) that were <50kb in size were used. There were 574 ‘downs’ and 672 ‘ups’ in this experiment. Each disclosure had one or more keywords and *possibly* one or more recognised FEP, named entity or type of financial object. As discussed in Sections 5.3.1 and 5.3.2, FEPs were not recognised in a significant number of disclosures in both S&P datasets. We conducted this experiment to see if a larger dataset changed the results in any way, and to see if the lack of FEPs in the majority of the disclosures improved or worsened the results.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 8.1% error rate on the training data and an average 47.9% error rate (or 52.1% accuracy rate) on the unseen test data (see the last two rows in Table 6.7). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

We also briefly examined the performance of the production rules derived from the original decision trees and found that, on average, there was a 34.8% error rate on the training data and an average 48.9% error rate (or 51.1% accuracy) on the unseen test data. If we wish to compare the average performance of the rules with the average performance of the trees, we must compare the former with the performance of the unpruned or original tree. The training error with the rules (34.8%) was significantly higher than the training error with the original tree (2.4%, see the second last row and third column in Table 6.7) and the error rate on the unseen test data was also higher, although not as significantly (48.9% vs. 46.8% respectively). In this experiment, the rules performed slightly worse on the unseen test data. The remainder of this section focuses on the output from the decision trees.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	665	28 (2.5%)	503	68 (6.1%)	(32.2%)
	<b>665</b>	<b>50 (40.0%)</b>	<b>503</b>	<b>50 (40.0%)</b>	<b>(32.2%)</b>
2	675	27 (2.4%)	433	98 (8.7%)	(31.9%)
	<b>675</b>	<b>60 (48.0%)</b>	<b>433</b>	<b>64 (51.2%)</b>	<b>(31.9%)</b>
3	693	24 (2.1%)	459	89 (7.9%)	(32.1%)
	<b>693</b>	<b>66 (52.8%)</b>	<b>459</b>	<b>69 (55.2%)</b>	<b>(32.1%)</b>
4	687	28 (2.5%)	427	104 (9.3%)	(32.1%)
	<b>687</b>	<b>58 (46.4%)</b>	<b>427</b>	<b>59 (47.2%)</b>	<b>(32.1%)</b>
5	669	23 (2.1%)	473	72 (6.4%)	(31.2%)
	<b>669</b>	<b>59 (47.2%)</b>	<b>473</b>	<b>58 (46.4%)</b>	<b>(31.2%)</b>
6	689	21 (1.9%)	465	90 (8.0%)	(32.6%)
	<b>689</b>	<b>57 (45.6%)</b>	<b>465</b>	<b>54 (43.2%)</b>	<b>(32.6%)</b>
7	663	30 (2.7%)	411	113 (10.1%)	(32.3%)
	<b>663</b>	<b>52 (41.9%)</b>	<b>411</b>	<b>55 (44.4%)</b>	<b>(32.3%)</b>
8	675	29 (2.6%)	471	89 (7.9%)	(32.6%)
	<b>675</b>	<b>60 (48.4%)</b>	<b>471</b>	<b>61 (49.2%)</b>	<b>(32.6%)</b>
9	687	32 (2.9%)	459	100 (8.9%)	(33.2%)
	<b>687</b>	<b>65 (52.4%)</b>	<b>459</b>	<b>66 (53.2%)</b>	<b>(33.2%)</b>
10	681	28 (2.5%)	481	88 (7.8%)	(33.0%)
	<b>681</b>	<b>56 (45.2%)</b>	<b>481</b>	<b>61 (49.2%)</b>	<b>(33.0%)</b>
<i>Train:</i>	678.4	27.0 (2.4%)	458.2	91.1 (8.1%)	(32.3%)
<i>Test:</i>	678.4	58.3 (46.8%)	458.2	59.7 (47.9%)	(32.3%)

Table 6.7: Number of errors and tree sizes for all cases in the 2005-2008 dataset: subset 3 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.8 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 67 test cases as UP (second row, last column) and 47 of these were correctly classified (second row, second column). C4.5 classified 58 cases as DOWN (second row, last column) and 28 were correctly classified (second row, third column). In the second and third columns, figures highlighted in bold are correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>47</b>	20	(a): class UP	67
	30	<b>28</b>	(b): class DOWN	58
<i>2</i>	<b>35</b>	32	(a): class UP	67
	32	<b>26</b>	(b): class DOWN	58
<i>3</i>	<b>32</b>	35	(a): class UP	67
	34	<b>24</b>	(b): class DOWN	58
<i>4</i>	<b>32</b>	35	(a): class UP	67
	24	<b>34</b>	(b): class DOWN	58
<i>5</i>	<b>35</b>	32	(a): class UP	67
	26	<b>32</b>	(b): class DOWN	58
<i>6</i>	<b>40</b>	27	(a): class UP	67
	27	<b>31</b>	(b): class DOWN	58
<i>7</i>	<b>37</b>	30	(a): class UP	67
	25	<b>32</b>	(b): class DOWN	57
<i>8</i>	<b>32</b>	35	(a): class UP	67
	26	<b>31</b>	(b): class DOWN	57
<i>9</i>	<b>32</b>	36	(a): class UP	68
	30	<b>26</b>	(b): class DOWN	56
<i>10</i>	<b>36</b>	32	(a): class UP	68
	29	<b>27</b>	(b): class DOWN	56
<i>Total</i>	<b>358</b>	314	(a): class UP	672
	283	<b>291</b>	(b): class DOWN	574

Table 6.8: Confusion matrix for the 2005-2008 dataset: subset 3 (C4.5).

If we add all the correct UP classifications, we find that 358/672 were correctly classified (53.3%). Similarly, if we add all the correct DOWN classifications, we find that 291/574 were correctly classified (50.7%). The next section analyses and discusses the results presented in Sections 6.2.2 to 6.2.5.

## 6.2.6 Analysis and Discussion: C4.5 Results

Looking at the error rates on training data for all four experiments, it ranged from 1.6% (subsets 1 and 2 combined) to 8.1% (subset 3), averaging at about 5.1%. These error rates are reasonably low, so we would hope to achieve good results with the test cases.

In the 1997-2000 dataset (Section 6.2.2), the average accuracy on ‘up’ and ‘down’ test cases was 52.1%. Likewise, in the 2005-2008 dataset (Section 6.2.3), the average accuracy was 52.5%. If we were to randomly assign each disclosure to the ‘up’ or ‘down’ category, we might expect a 50-50 result (see, for example, Kryzanowski et al 1993). Using this very crude measure, we can see that our recogniser and classifiers yielded slightly better-than-chance results, for both datasets.

However, on closer examination of the ‘ups’ and ‘downs’, we can see that the classifiers correctly classified 53.3% of the ‘ups’ and 50.6% of the ‘downs’ (subset 1) and 63.2% of the ‘ups’ and 37.6% of the ‘downs’ (subset 2). However, as there is not an even number of ‘ups’ and ‘downs’ in either dataset, we should not compare these figures to 50%. Ideally, we would undertake a hypothesis test for a population proportion to prove that these results are statistically significant but as we do not know the true proportion of ‘ups’ and ‘downs’ in the entire disclosure universe, we cannot compute this value. An alternative method of evaluating the performance is to compute the proportion of ‘ups’ and ‘downs’ in our sample dataset and use those values as benchmarks when evaluating the performance of the classifiers.

Looking at the ‘ups’ in the 1997-2000 dataset, there are 90 ‘ups’ in the dataset of 169 disclosures (53.3%). Coincidentally, this figure matches the percentage of correctly classified ‘ups’. If we were to arbitrarily classify every document as ‘up’, we would be correct 53.3% of the time, so our prototype recogniser and classifiers does not yield improved results over that arbitrary method—in fact, it achieves the same result. However, if we compute similar values for the ‘downs’, we find that the dataset comprises 46.7% ‘downs’, but our prototype system correctly classifies 50.6% of them, so this demonstrates that our system is better than arbitrary classification.

One possible explanation for these results for the 1997-2000 dataset could be that corporations were only required to file certain types of events prior to 2004 and there was greater flexibility with regards to the filing deadline (see Section 4.2 for a discussion). At the time, the controls were not as strict as they are today, so corporations did not have to file as many types of negative news events and would only have mentioned negative news if they really had to do so; consequently, the classifiers may have found it relatively easy to identify negative news when it did occur. Also, in Section 5.3.3 we revealed that there were proportionately more words in the ‘ups’, compared to the ‘downs’. This could partially explain the difficulties faced by the classifiers when attempting to classify the ‘ups’ relative to the ‘downs’. This seems to correlate with early findings by Hildebrandt and Snyder (1981), who found that there were significantly more occurrences of positive words than neutral or negative words, regardless of whether it was a financially good or bad year (see Section 2.3 for a discussion). As a result, positive news may be more difficult to classify than negative news.

Looking at the ‘ups’ in the 2005-2008 dataset, there are 163 ‘ups’ in the dataset of 280 disclosures (58.2%). If we were to arbitrarily classify every document as ‘up’, we would be correct 58.2% of the time. However, our recogniser and classifiers correctly classified the ‘ups’ 63.2% of the time, so our system outperforms the arbitrary method. If we compute similar values for the ‘downs’, however, our system only correctly classifies 37.6% of them, whereas arbitrary classification would yield 41.8% accuracy.

Post 2004, the ‘ups’ appeared to be easier to classify than the ‘downs’, which is a reversal on the previous period. A preliminary pattern analysis (see Section 5.3.3) revealed that the ‘ups’ became more concise (relative to the ‘downs’) during this period; one possible reason for this improved conciseness could be the stricter filing regulations, which were concerned with content scope and filing deadlines. As the filing deadlines are shorter nowadays, we might expect disclosures to have a greater impact on share price return than previously, so there is more at stake for the corporation. Also, as corporations are now required to file a greater number of events, they may adopt strategies which disguise or frame important negative news in

the midst of not-so-important positive news (Loughran and McDonald 2011b). This noisy “positive” news could make negative news more difficult to classify. In an informal poll undertaken by Loughran and McDonald (2011a), a small number of accounting firms admitted that they would bury an “awkward revelation in an overwhelming amount of uninformative text and data” (p.2). As discussed in Section 2.4, Tetlock et al (2008) found that reaction to negative information usually takes place within one day, so perhaps the reaction to this news had already been absorbed into prices. This offers another possible reason for the poor classification of ‘downs’.

In Section 6.2.4, we discussed the results when we merged the 1997-2000 and 2005-2008 datasets. We found that the average accuracy was 52.8% for ‘ups’ and ‘downs’. Using the random classification method, our result is marginally better than the 50-50 result. However, on closer examination of the ‘ups’ and ‘downs’, we can see that the classifiers correctly classified 56.9% of the ‘ups’ and 47.4% of the ‘downs’. If we compare these figures to the actual proportion of ‘ups’ and ‘downs’ in this combined dataset (56.3% and 43.7%), we can see that the classifiers were marginally better with the ‘ups’ but better still with the ‘downs’. We must remember that each of these disclosures had one or more recognised FEP and all had a number of keywords and/or types of financial object. However, as this merged dataset comprised data that was filed pre- and post- the 2004 rule changes, it is difficult to determine the extent to which the SEC rule changes may have impacted these results.

In Section 6.2.5, we presented the results for 1,246 8-Ks filed for the 50 S&P500 companies, during 2005-2008, regardless of whether FEPs were recognised in those disclosures or not. We should remember here that only disclosures (cases) that were <50kb in size were used, but other than that, each case had one or more keywords and *possibly* one or more FEP, name entity or type of financial object. In this experiment, we found that the average accuracy was 52.1% for ‘ups’ and ‘downs’. On closer examination of the ‘ups’ and ‘downs’, we can see that the classifiers correctly classified 53.3% of the ‘ups’ and 50.7% of the ‘downs’. If we compare these figures to the actual proportion of ‘ups’ and ‘downs’ in this relatively large dataset (53.9% and 46.1%), we can see that our classifiers performed marginally worse with the ‘ups’ but better with the ‘downs’.

In Chapter 5, we discussed the FEP types that our recogniser found in each dataset. At most, 18 unique FEP types were identified by the recogniser (see Tables 5.3, 5.11, and 5.19). Of these 18 FEP types, 15 were used by one or more decision trees in the various C4.5 experiments (see Appendix 26 for a listing). However, we should point out at this stage that the experiment discussed in Section 6.2.2 (subset 1) only used two FEP types in the decision trees and the experiment discussed in Section 6.3.3 only used one FEP type (subset 2). One possible reason for this could be the relatively small size of both datasets (169 and 280 disclosures respectively); it is likely that there were too few cases for the very large number of features (1,635). The experiment discussed in Section 6.2.4 used nine FEP types (449 disclosures, subsets 1 and 2 combined) in the decision trees and the experiment discussed in Section 6.2.5 used 15 FEP types (1,246 disclosures, subset 3). The increasing number of FEP types recognised in each experiment suggests that larger datasets find the FEPs more useful. Even though only 280 of these 1,246 cases had one or more recognised FEP (all had one or more keywords, named entities, and/or types of financial object), it could be that the absence of FEPs in the other cases was a useful predictive attribute. Nonetheless, we need to be careful not to assume that this is the case, as the recall of our prototype recogniser still needs further work. Ideally, every disclosure or case should have at least one recognised FEP—this is currently not the case.

To summarise, if we compare the actual classification accuracy for ‘ups’ and ‘downs’ separately, with the actual proportion of ‘ups’ and ‘downs’ in each experiment, the results for C4.5 using FEPs and keywords are as follows:

*1997-2000 dataset (subset 1):*

- C4.5: ‘ups’ **equals** the arbitrary classification method.
- C4.5: ‘downs’ **outperforms** arbitrary classification.

*2005-2008 dataset (subset 2):*

- C4.5: ‘ups’ **outperforms** arbitrary classification.
- C4.5: ‘downs’ does not outperform arbitrary classification.

*1997-2000 and 2005-2008 datasets (subsets 1 and 2):*

- C4.5: ‘ups’ **marginally outperforms** arbitrary classification.
- C4.5: ‘downs’ **outperforms** arbitrary classification.

*2005-2008 dataset (subset 3):*

- C4.5: ‘ups’ does not outperform arbitrary classification.
- C4.5: ‘downs’ **outperforms** arbitrary classification.

Summarising these results, we can say that in three of the four experiments, C4.5 equals or outperforms the arbitrary classification method with the ‘ups’. Also, in three of the four experiments, C4.5 outperforms arbitrary classification with the ‘downs’. We should also note here that the C4.5 training error rates were quite low, ranging from 1.6% to 8.1% (see Sections 6.2.2 to 6.2.5). This low training error would have helped the classification of test cases.

In Section 6.4, we will present experiments which used C4.5 and n-grams, rather than FEPs and keywords, for classification. These n-gram experiments were undertaken as a baseline approach, to demonstrate the value (if any) of using FEPs for classification. In Chapter 7, we will discuss the limitations of our datasets and our FEP recogniser and how these factors may have impacted the overall results.

## 6.3 Support Vector Machine Classification using Financial Event Phrases and Keywords

In this section, we first present an overview of support vector machines (SVMs) in general and then focus specifically on one implementation of SVMs, known as SVM-Light. We then present the results from a number of SVM-Light experiments in Sections 6.3.2 to 6.3.5:

- Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light).
- Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light).
- Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM-Light).
- Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light).

In Section 6.3.6, we will analyse and discuss the results from all four experiments.

### 6.3.1 Background to Support Vector Machines and SVM-Light

Support Vector Machines are a type of Kernel Method (KM) (Bennett and Campbell 2000; Berthold and Hand 2003). KMs are becoming increasingly popular as they are regarded as simple and computationally efficient (like linear algorithms), flexible (like non-linear systems) and they comprise rigorous statistical approaches which avoid common problems like overfitting<sup>40</sup> (Berthold and Hand 2003). For these reasons, KMs such as SVM are popular in fields such as bioinformatics and text classification (Cristianini and Shawe-Taylor 2000; Bennett and Campbell 2000; Berthold and Hand 2003; Fradkin and Muchnik 2006). We decided to experiment with SVMs because they have also been used with various levels of success in the financial domain (see, for example, Schumaker and Chen 2006, Mittermayer and Knolmayer 2006a, Fung et al 2005, Antweiler and Frank 2004, Koppel and Shtrimberg 2004, and van Bunningen 2004 in Chapter 3). SVMs have also been used quite successfully in time series forecasting, which is beyond the scope of our research (but see, for example, Tay et al 2003).

---

<sup>40</sup> Overfitting is said to occur when false relations are detected that are “the effect of chance and do not reflect any underlying property of the data” (Berthold and Hand, 2003, pp.176-177).

In an early paper, Cortes and Vapnik (1995) defined a support vector network as follows:

*“The support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space  $Z$  through some non-linear mapping chosen a priori. In this space a linear decision surface is constructed with special properties that ensure high generalization ability of the network”* (p.274).

Some of the major advantages of SVMs is that they are able to optimise generalisation bounds (i.e. minimise the upper bound of the generalisation error) and they are capable of dealing with hundreds of thousands of examples (Joachims 1998; Berthold and Hand 2003; Tay et al 2003). For an overview of how SVMs can be applied to text categorisation in general, see Joachims (1998).

Different types of SVM exist, including the maximal margin classifier and the soft margin optimisation classifier. The maximal margin classifier is the simplest form, but it is not very useful for real-world problems especially if there is no linear separation in the feature space (i.e. the data is noisy). The soft margin optimisation classifier, on the other hand, is more complicated to understand but it has more robust bounds that are capable of dealing with noise and outliers (Berthold and Hand 2003).

We chose Joachims’s SVM-Light as it is “one of the most widely used SVM classification and regression package [sic]” (p.387). SVM-Light is an implementation of Vapnik’s SVMs (Joachims 1998).

SVM-Light only requires two data files—a training file and a test file (see Figures 6.4 and 6.5 respectively). On the first line in Figure 6.4, the -1 specifies the class for the first case (-1 equates to ‘down’). On the second line, 1 specifies the class for the second case (1 equates to ‘up’). For each case in the training file, each feature is assigned a value e.g. feature 1 has a value of ‘TRUE’ which means that the feature applies to that case and feature 2 has a value of ‘FALSE’, so that feature does *not* apply to this case. The values for each feature are presented one after the other. The

training and test files both take the same format. Unlike C4.5, there is no need for a separate .names file, as the features are specified in the training and test files, along with the values.

```
-1 1:TRUE 2:FALSE 3:FALSE,...,1635:TRUE  
1 1:FALSE 2:TRUE 3:TRUE,...,1635:FALSE  
...
```

Figure 6.4: Sample content for the SVM-Light .data training file (two cases).

```
-1 1:TRUE 2:TRUE 3:FALSE,...,1635:FALSE  
1 1:TRUE 2:FALSE 3:TRUE,...,1635:FALSE  
...
```

Figure 6.5: Sample content for the SVM-Light .test test file (two cases).

In Section 6.2.1, we discussed the benefits of using K-fold cross-validation with C4.5. SVM-Light provides for another type of cross-validation, known as leave-one-out estimation (Joachims 2000). With this method, one sample is removed and all others are used for training, leading to a classification rule. This rule is tested on the one hold out case and this process is then repeated for all the training examples. The number of misclassifications divided by the number of cases gives the leave-one-out estimate of the generalisation error. Essentially, this method considers the impact that a single training example can have on the performance of the learner but this is very small for most practical problems (Joachims 2000).

As we wanted to compare the results of the C4.5 experiments with the results from the SVM-Light package, we decided instead to partition the data into 10 fairly equal-sized partitions. We then tested on a different partition for each of the ten folds, using the remaining nine partitions as training data. To create the ten partitions, we randomly generated 20% testing data and 80% training data ten times, from all the available data. This approach is not identical to the K-fold cross validation approach used by C4.5 (and discussed in Section 6.2), but it does mean that different combinations of training and test data were used at each fold. Whilst this workaround does *not* guarantee that each case appears in only one test set, we think that averaging

the results over ten folds should still be a good predictor of the model's error rate.

SVM-Light comprises a learning module (`svm_learn`) and a classification module (`svm_classify`). To use SVM-light, one must first learn a classification model from the training data file using `svm_learn`. This model is then applied to the test data file using `svm_classify`. SVM-Light output differs from C4.5 output in that it does not report on the classification errors for the two classes separately—it only presents the accuracy for both together. To counteract this, we decided to train the ‘ups’ and ‘downs’ together to generate the model but we tested the ‘ups’ and ‘downs’ separately. Again, this was a workaround, but we wanted to make the experiments as comparable as possible. The model generated by the classifier presents some preliminary details including the number of features, the number of training documents, and the number of support vectors. Each support vector is then presented, preceded with the alpha or weights for that vector. The model is then applied to the training data and the number of training misclassifications is calculated. The output on the training data also specifies the number of support vectors and the Xi-Alpha-estimates of the error, recall, and precision. These Xi-Alpha-estimates are conservatively biased (see Joachims 2000 for a discussion of these and other generalisation errors applicable to SVMs). The output from the test cases is fairly straightforward—SVM-Light presents the accuracy on the test set in terms of how many cases there were and how many were correctly and incorrectly classified.

In Sections 6.3.2 to 6.3.5, we present the results from our SVM-Light experiments and in Section 6.3.6 we analyse and discuss the results.

### **6.3.2 Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light)**

This experiment used 169 disclosures from the 1997-2000 dataset and comprised 90 ‘ups’ and 79 ‘downs’. Each of these disclosures had one or more recognised FEPs within them, according to our recogniser. We used 1,635 attributes for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects. This is the same data that was described in Section 6.2.2, for the C4.5 algorithm. As discussed in Section 5.5, we used keywords as well as FEPs, because the overall FEP recognition performance was poor.

Using our method of ten-fold cross validation (see Section 6.3.1), SVM-Light had an average 20% error rate on the training data and an average 50.3% error rate (or 49.7% accuracy rate) on the unseen test data (see the last row and column in Table 6.9). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 30.0% of the ‘ups’ (70% accuracy) but 70.6% of the ‘downs’ (29.4% accuracy).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	136	23 (16.9%)	<=94.85%	18 (u) 15 (d)	7 (38.9%) 8 (53.3%)
2	135	28 (20.7%)	<=92.59%	18 (u) 16 (d)	4 (22.2%) 12 (75.0%)
3	135	41 (30.4%)	<=94.07%	18 (u) 16 (d)	1 (5.6%) 13 (81.3%)
4	135	23 (17.0%)	<=91.85%	18 (u) 16 (d)	4 (22.2%) 15 (93.8%)
5	135	26 (19.3%)	<=89.63%	18 (u) 16 (d)	6 (33.3%) 12 (75.0%)
6	135	28 (20.7%)	<=93.33%	18 (u) 16 (d)	6 (33.3%) 13 (81.3%)
7	135	19 (14.1%)	<=94.07%	18 (u) 16 (d)	10 (55.6%) 7 (43.8%)
8	135	33 (24.4%)	<=94.07%	18 (u) 16 (d)	5 (27.8%) 13 (81.3%)
9	135	25 (18.5%)	<=94.07%	18 (u) 16 (d)	7 (38.9%) 9 (56.3%)
10	135	28 (20.7%)	<=93.33%	18 (u) 16 (d)	4 (22.2%) 11 (68.8%)
<i>Average</i>	135	27 (20.0%)	<=93.18%	18 (u) 16 (d)	5.4 (30.0%) 11.3 (70.6%)

Table 6.9: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 dataset: subset 1 (SVM-Light).

### 6.3.3 Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light)

This experiment used 280 disclosures from the 2005-2008 dataset, comprising 163 ‘ups’ and 117 ‘downs’. Each of these disclosures had one or more recognised FEPs within them, according to our recogniser. Like the previous experiment, we used 1,635 attributes for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects.

Using our method of ten-fold cross validation, SVM-Light had an average 35.3% error rate on the training data and an average 50.4% error rate (or 49.6% accuracy rate) on the unseen test data (see the last row and column in Table 6.10). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 1.2% of the ‘ups’ (98.8% accuracy) but 99.6% of the ‘downs’ (0.4% accuracy), demonstrating a huge difference in accuracy levels. This seems to suggest that SVM-Light behaved like a default classifier here, assigning the majority of test cases to the class with the most data (Joachims 2000).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	224	80 (35.7%)	$\leq 85.27\%$	33 (u) 23 (d)	0 (0%) 22 (95.7%)
2	224	81 (36.2%)	$\leq 85.71\%$	33 (u) 23 (d)	0 (0%) 23 (100%)
3	223	85 (38.1%)	$\leq 84.30\%$	33 (u) 24 (d)	0 (0%) 24 (100%)
4	224	82 (36.6%)	$\leq 87.50\%$	33 (u) 23 (d)	0 (0%) 23 (100%)
5	223	77 (34.5%)	$\leq 84.75\%$	33 (u) 24 (d)	0 (0%) 24 (100%)
6	225	75 (33.3%)	$\leq 84.89\%$	32 (u) 23 (d)	1 (u) – 3.1% 23 (100%)
7	223	72 (32.3%)	$\leq 85.65\%$	33 (u) 24 (d)	2 (u) – 6.1% 24 (100%)
8	224	79 (35.3%)	$\leq 88.39\%$	33 (u) 23 (d)	1 (u) – 3.0% 23 (100%)
9	224	81 (36.2%)	$\leq 87.95\%$	33 (u) 23 (d)	0 (0%) 23 (100%)
10	224	78 (34.8%)	$\leq 86.16\%$	33 (u) 23 (d)	0 (0%) 23 (100%)
<i>Average</i>	224	79 (35.3%)	$\leq 86.06\%$	33 (u) 23 (d)	0.4 (1.2%) 23 (99.6%)

Table 6.10: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 2 (SVM-Light).

#### 6.3.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM-Light).

This experiment used all the data from the previous two experiments (Sections 6.3.2. and 6.3.3). There were 449 disclosures, comprising 253 ‘ups’ and 196 ‘downs’. Each of these disclosures had one or more recognised FEPs within them, according to

our recogniser. Like the previous two experiments, we used 1,635 attributes for classification, comprising 1,568 hand-chosen keywords, 49 types of FEP, and 18 types of named entities and financial objects.

Using our method of ten-fold cross validation, SVM-Light had an average 30.6% error rate on the training data and an average 52.7% error rate (or 47.3% accuracy rate) on the unseen test data (see the last row and column in Table 6.11). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 12.7% of the ‘ups’ (87.3% accuracy) but 92.6% of the ‘downs’ (7.4% accuracy), demonstrating a huge difference in accuracy levels. Again, it seems as if SVM-Light behaved like a default classifier here.

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	359	117 (32.6%)	$\leq 88.86\%$	51 (u) 39 (d)	8 (15.7%) 36 (92.3%)
2	359	100 (27.9%)	$\leq 88.30\%$	51 (u) 39 (d)	8 (15.7%) 35 (89.7%)
3	359	113 (31.5%)	$\leq 88.86\%$	51 (u) 39 (d)	0 (0%) 38 (97.4%)
4	358	112 (31.3%)	$\leq 87.99\%$	51 (u) 40 (d)	3 (5.9%) 38 (95.0%)
5	358	112 (31.3%)	$\leq 89.39\%$	51 (u) 40 (d)	9 (17.6%) 37 (92.5%)
6	359	111 (30.9%)	$\leq 87.47\%$	51 (u) 39 (d)	14 (27.5%) 33 (84.6%)
7	359	113 (31.5%)	$\leq 88.30\%$	51 (u) 39 (d)	3 (5.9%) 36 (92.3%)
8	359	102 (28.4%)	$\leq 88.30\%$	51 (u) 39 (d)	10 (19.6%) 33 (84.6%)
9	359	107 (29.8%)	$\leq 88.58\%$	51 (u) 39 (d)	7 (13.7%) 38 (97.4%)
10	359	114 (31.8%)	$\leq 88.02\%$	51 (u) 39 (d)	3 (5.9%) 37 (94.9%)
<i>Average</i>	359	110 (30.6%)	$\leq 88.41\%$	51 (u) 39 (d)	6.5 (12.7%) 36.1 (92.6%)

Table 6.11: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (SVM-Light).

### 6.3.5 Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light)

This experiment used 1,246 disclosures that were filed for the 50 S&P500 companies, during the period 2005-2008. Only the disclosures that were <50kb in size were used. Each disclosure had one or more keywords and *possibly* one or more recognised FEP, named entity or type of financial object. As discussed in Section 5.3.2, a significant number of cases did not have any recognised FEPs. We conducted this experiment to see if a larger dataset changed the results in any way, and to see if the lack of FEPs in the majority of the disclosures improved or worsened the results.

Using our method of ten-fold cross validation, SVM-Light had an average 33.0% error rate on the training data and an average 49.0% error rate (or 51.0% accuracy rate) on the unseen test data (see the last row and column in Table 6.12). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 12.2% of the ‘ups’ (87.8% accuracy) but 85.7% of the ‘downs’ (14.3% accuracy), demonstrating a huge difference in accuracy levels.

Partition	# Training cases	# Training cases misclassified	XiAlpha-estimate of the error	# Up (u) and down (d) test cases	# Incorrect up and down classifications
1	996	339 (34.0%)	<=93.17%	135 (u) 115 (d)	11 (8.1%) 97 (84.3%)
2	996	341 (34.2%)	<=92.47%	135 (u) 115 (d)	23 (17.0%) 98 (85.2%)
3	996	349 (35.0%)	<=93.17%	135 (u) 115 (d)	11 (8.1%) 99 (86.1%)
4	996	304 (30.5%)	<=93.78%	135 (u) 115 (d)	20 (14.8%) 92 (80.0%)
5	996	350 (35.1%)	<=93.27%	135 (u) 115 (d)	16 (11.9%) 100 (87.0%)
6	996	321 (32.2%)	<=93.07%	135 (u) 115 (d)	25 (18.5%) 95 (82.6%)
7	997	329 (33.0%)	<=93.38%	135 (u) 114 (d)	10 (7.4%) 104 (91.2%)
8	996	349 (35.0%)	<=92.87%	135 (u) 115 (d)	12 (8.9%) 104 (90.4%)
9	997	340 (34.1%)	<=93.18%	134 (u) 115 (d)	14 (10.4%) 98 (85.2%)
10	996	319 (32.0%)	<=92.87%	135 (u) 115 (d)	23 (17.0%) 99 (86.1%)
Average	996	334 (33.0%)	<=93.11%	135 (u) 115 (d)	16.5 (12.2%) 98.6 (85.7%)

Table 6.12: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subsets 3(SVM-Light).

### 6.3.6 Analysis and Discussion: SVM-Light Results

Looking at the error rates on training data for all four experiments, it ranged from 20.0% (subset 1) to 35.3% (subset 2), averaging at about 30.0%. Ideally, these error rates need to be significantly lower, if we want to secure low error rates on test data. In Section 6.5, we will present experiments which used SVM-Light and n-grams, rather than FEPs and keywords, for classification. These additional experiments were undertaken as a baseline approach, to determine the value (if any) of using FEPs for classification.

In the 1997-2000 experiment (Section 6.3.2), the average accuracy on ‘up’ and ‘down’ test cases was 49.7%. In the 2005-2008 experiment (Section 6.3.3), the average accuracy was 49.6%. If we were to randomly assign each disclosure to the ‘up’ or ‘down’ category, we might expect a 50-50 result. Using this very crude measure, we can see that our recogniser and SVM-Light yielded slightly lower results, for both datasets. The average error rate over all four experiments was 50.6%, just marginally above chance.

Looking more closely at the ‘ups’ and ‘downs’ separately for the first two experiments, we can see that SVM-Light correctly classified 70.0% of the ‘ups’ but only 29.4% of the ‘downs’ (subset 1) and 98.8% of the ‘ups’ and 0.4% of the ‘downs’ (subset 2). As discussed in Section 6.2.6 (C4.5 analysis and discussion), there were an uneven number of ‘ups’ and ‘downs’ in each experiment, so an alternative method of evaluating the performance is to compute the proportion of ‘ups’ and ‘downs’ in our sample dataset and use those values as benchmarks when evaluating the performance of both systems, rather than random assignment. If we were to arbitrarily classify all cases as ‘up’ or ‘down’, we would hope that our system’s accuracy outperforms the actual proportion of ‘ups’ or ‘downs’.

Looking first at the ‘ups’ in the 1997-2000 dataset (subset 1), there were 90 ‘ups’ in the dataset of 169 disclosures (53.3%). SVM-Light correctly classified 70.0% of the ‘ups’, so our prototype recogniser and SVM-Light yielded improved results over the arbitrary classification method. However, if we compute similar values for the ‘downs’, we find that the dataset comprises 46.7% ‘downs’, but our system only

correctly classified 29.4% of them. As discussed in Section 4.2, these cases were filed before the 2004 rule changes, so the filing deadlines were not as tight as they are today and companies had greater flexibility regarding the types of events they had to file. Unlike the C4.5 results for this period (Section 6.2.2), the ‘ups’ were easier to classify than the ‘downs’.

Looking at the ‘ups’ in the 2005-2008 dataset (subset 2), there were 163 ‘ups’ in the dataset of 280 disclosures (58.2%). If we were to arbitrarily classify every document as ‘up’, we would be correct 58.2% of the time. However, our system correctly classified 98.8% of the ‘ups’, so it far outperformed the arbitrary method<sup>41</sup>. If we compute similar values for ‘downs’, our system only correctly classified 0.4%, whereas arbitrary classification would yield 41.8% accuracy. We know that this period was post the 2004 rule changes, so the filing requirements were much stricter. We also know from Section 5.3.3, that the ‘ups’ became more precise (relative to the ‘downs’) during this period, compared to the 1997-2000 period, so the ‘ups’ possibly contained less ‘noise’ than before and were therefore even easier to classify. The system performed very poorly with the ‘downs’, which could possibly have been caused by strategies to disguise negative news in the midst of not-so-important positive news. For example, in Section 2.3, we discussed one study that identified a shift away from the writer of the message when disclosing negative news in annual reports (Thomas 1997). Thomas noted that references to negative news (caused by nonhuman participants) were often followed up with suggestions about how the situation might improve—words such as *nevertheless* and *however* were often used to lessen the impact. She also outlined how the use of certain words can mean different things depending on the context. For example, if *profitability* is mentioned, does this mean the company is now (more) profitable (compared to previously) or that the company *may* return to profitability at some stage in the future?

---

<sup>41</sup> As discussed in Section 6.3.3, we believe SVM behaved like a default classifier here, assigning the majority of cases to the class with the most data. Joachims (2000) encountered some similar findings with a small training dataset of Reuters news categories. Interestingly, though, in our experiments, SVM did not appear to use the default class with the 1997-2000 dataset, which was even smaller.

In Section 6.3.4, we discussed the SVM-Light findings when we merged the 1997-2000 and 2005-2008 datasets (subsets 1 and 2). We found that the average accuracy was 47.3% for the ‘ups’ and ‘downs’. This result would not be as good as the random classification (50-50) method. On closer examination of the ‘ups’ and ‘downs’, we can see that SVM-Light correctly classified 87.3% of the ‘ups’ but only 7.4% of the ‘downs’—these results are respectively better and lower than the actual portion of ‘ups’ and ‘downs’ in the combined dataset (56.3% and 43.7% respectively). As discussed previously, this dataset comprises data filed before *and* after the 2004 rule changes, so the content requirements and filing deadlines varied greatly in these disclosures. Therefore, it is difficult to tell if the 2004 rule changes impacted the results.

In Section 6.3.5, we presented the results for the 1,246 8-Ks filed for the 50 S&P500 companies during 2005-2008, regardless of whether our prototype recogniser identified FEPs are not. Each of these cases had one or more keywords and *possibly* one or more recognised FEPs, named entities or terms. The only other criterion was that all cases were <50kb in size. In this experiment, SVM-Light correctly classified 87.8% of the ‘ups’ but only 14.3% of the ‘downs’—these figures can be compared to the actual proportion of ‘ups’ and ‘downs’ (53.9% and 46.1%). Once again, our prototype recogniser and SVM-Light outperformed the arbitrary classification method for the ‘ups’, but not for the ‘downs’.

To summarise, if we compare the actual classification accuracy for the ‘ups’ and ‘downs’ separately, with the actual proportion of ‘ups’ and ‘downs’ in each experiment, the results for SVM-Light using FEPs and keywords are as follows:

*1997-2000 dataset (subset 1):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*2005-2008 dataset (subset 2):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*1997-2000 and 2005-2008 datasets (subsets 1 and 2):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*2005-2008 dataset (subset 3):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

Summarising these results, we can say that SVM-Light always outperforms arbitrary classification with the ‘ups’ but it never outperforms arbitrary classification with the ‘downs’. We must remember though that SVM-Light behaved like a default classifier, using the largest class (‘up’) as the default class. We should also note here that the SVM-Light training error rates were quite high, ranging from 20% to 35.3% (see Sections 6.3.2 to 6.3.5).

## 6.4 Decision Tree Classification using N-Grams

In Section 3.2, we discussed various document representations, including single keywords, phrases, terms, named entities, and n-grams. In Section 6.2 we presented an overview of classification using decision trees, with a specific focus on the C4.5 suite of programs. We also presented four C4.5 experiments which used FEPs and keywords as document features. In this section, we present C4.5 experiments which used automatically-extracted *n-grams* (where  $n=5$ ) as document features.

For each of these experiments, we generated a list of the top 1,000 most-frequently occurring five-word n-grams for the disclosures in each dataset. Using the relevant n-gram lists, we then automatically extracted the five-word n-grams that applied to each disclosure.

In Sections 6.4.1 to 6.4.4, we present the results from each of these experiments:

- Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5)
- Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5)
- Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5)
- Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5)

In Section 6.4.5, we will analyse and discuss the results from all four experiments.

### 6.4.1 Classification Results for the 1997-2000 Dataset: Subset 1 (C4.5)

This experiment used the 169 disclosures from the 1997-2000 dataset that were also used in the FEP and keyword experiments with C4.5 and SVM-Light (see Sections 6.2.2 and 6.3.2 respectively). The dataset comprised 90 ‘ups’ and 79 ‘downs’, each with one or more automatically extracted five-word n-grams.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 17.9% error rate on the training data and an average 46.7% error rate (or 53.3% accuracy rate) on the unseen test data (see the last two rows in Table 6.13). The table also shows the number of errors for training and test data, for each fold. For each

fold, the rows in bold type signify the evaluations on unseen test cases.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	89	18 (11.8%)	41	32 (21.1%)	(38.5%)
	<b>89</b>	<b>6 (35.3%)</b>	<b>41</b>	<b>8 (47.1%)</b>	<b>(38.5%)</b>
2	83	13 (8.6%)	39	28 (18.4%)	(35.2%)
	<b>83</b>	<b>10 (58.8%)</b>	<b>39</b>	<b>11 (64.7%)</b>	<b>(35.2%)</b>
3	85	14 (9.2%)	65	17 (11.2%)	(36.7%)
	<b>85</b>	<b>6 (35.3%)</b>	<b>65</b>	<b>5 (29.4%)</b>	<b>(36.7%)</b>
4	85	14 (9.2%)	37	30 (19.7%)	(36.1%)
	<b>85</b>	<b>7 (41.2%)</b>	<b>37</b>	<b>9 (52.9%)</b>	<b>(36.1%)</b>
5	93	11 (7.2%)	29	32 (21.1%)	(34.5%)
	<b>93</b>	<b>10 (58.8%)</b>	<b>29</b>	<b>9 (52.9%)</b>	<b>(34.5%)</b>
6	81	17 (11.2%)	39	34 (22.4%)	(39.1%)
	<b>81</b>	<b>6 (35.3%)</b>	<b>39</b>	<b>8 (47.1%)</b>	<b>(39.1%)</b>
7	89	11 (7.2%)	43	24 (15.8%)	(34.3%)
	<b>89</b>	<b>8 (47.1%)</b>	<b>43</b>	<b>8 (47.1%)</b>	<b>(34.3%)</b>
8	85	13 (8.6%)	41	25 (16.4%)	(34.2%)
	<b>85</b>	<b>7 (41.2%)</b>	<b>41</b>	<b>8 (47.1%)</b>	<b>(34.2%)</b>
9	87	14 (9.2%)	47	26 (17.1%)	(36.7%)
	<b>87</b>	<b>6 (35.3%)</b>	<b>47</b>	<b>7 (41.2%)</b>	<b>(36.7%)</b>
10	93	13 (8.5%)	51	24 (15.7%)	(36.3%)
	<b>93</b>	<b>7 (43.8%)</b>	<b>51</b>	<b>6 (37.5%)</b>	<b>(36.3%)</b>
<i>Train (average):</i>	87.0	13.8 (9.1%)	43.2	27.2 (17.9%)	(36.2%)
<i>Test (average):</i>	87.0	7.3 (43.2%)	43.2	7.9 (46.7%)	(36.2%)

Table 6.13: Number of errors and tree sizes for the 1997-2000 dataset: subset 1 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.14 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 9 test cases as UP (second row, last column) but only 3 of these were correctly classified

(second row, second column). C4.5 classified 8 cases as DOWN (second row, last column) and 6 of these were correctly classified (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>3</b>	6	(a): class UP	9
	2	<b>6</b>	(b): class DOWN	8
<i>2</i>	<b>2</b>	7	(a): class UP	9
	4	<b>4</b>	(b): class DOWN	8
<i>3</i>	<b>6</b>	3	(a): class UP	9
	2	<b>6</b>	(b): class DOWN	8
<i>4</i>	<b>6</b>	3	(a): class UP	9
	6	<b>2</b>	(b): class DOWN	8
<i>5</i>	<b>3</b>	6	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>6</i>	<b>3</b>	6	(a): class UP	9
	2	<b>6</b>	(b): class DOWN	8
<i>7</i>	<b>4</b>	5	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>8</i>	<b>4</b>	5	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>9</i>	<b>5</b>	4	(a): class UP	9
	3	<b>5</b>	(b): class DOWN	8
<i>10</i>	<b>7</b>	2	(a): class UP	9
	4	<b>3</b>	(b): class DOWN	7
<i>Total</i>	<b>43</b>	47	(a): class UP	90
	32	<b>47</b>	(b): class DOWN	79

Table 6.14: Confusion matrix for the 1997-2000 dataset: subset 1 (C4.5).

If we add all the correct UP classifications, we find that 43/90 UP classifications were correctly classified (47.8%). If we compute a similar value for the DOWNS, we find that 47/79 (59.5%) were correctly classified. A discussion of these results can be found in Section 6.4.5.

### 6.4.2 Classification Results for the 2005-2008 Dataset: Subset 2 (C4.5)

This experiment used the 280 disclosures from the 2005-2008 dataset that were also used in the FEP and keyword experiments with C4.5 and SVM-Light (see Sections 6.2.3 and 6.3.3 respectively). The dataset comprised 163 ‘ups’ and 117 ‘downs’, each with one or more automatically extracted five-word n-grams.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 24.4% error rate on the training data and an average 49.3% error rate (or 50.7% accuracy rate) on the unseen test data (see the last two rows in Table 6.15). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	93	42 (16.7%)	35	65 (25.8%)	(36.9%)
	<b>93</b>	<b>16 (57.1%)</b>	<b>35</b>	<b>14 (50.0%)</b>	<b>(36.9%)</b>
2	107	42 (16.7%)	43	66 (26.2%)	(39.4%)
	<b>107</b>	<b>11 (39.3%)</b>	<b>43</b>	<b>16 (57.1%)</b>	<b>(39.4%)</b>
3	101	36 (14.3%)	63	49 (19.4%)	(35.9%)
	<b>101</b>	<b>17 (60.7%)</b>	<b>63</b>	<b>17 (60.7%)</b>	<b>(35.9%)</b>
4	109	46 (18.3%)	27	75 (29.8%)	(39.1%)
	<b>109</b>	<b>12 (42.9%)</b>	<b>27</b>	<b>15 (53.6%)</b>	<b>(39.1%)</b>
5	103	46 (18.3%)	45	67 (26.6%)	(39.3%)
	<b>103</b>	<b>12 (42.9%)</b>	<b>45</b>	<b>12 (42.9%)</b>	<b>(39.3%)</b>
6	119	42 (16.7%)	39	64 (25.4%)	(37.7%)
	<b>119</b>	<b>42 (16.7%)</b>	<b>39</b>	<b>16 (57.1%)</b>	<b>(37.7%)</b>
7	105	43 (17.1%)	33	67 (26.6%)	(37.9%)
	<b>105</b>	<b>13 (46.4%)</b>	<b>33</b>	<b>13 (46.4%)</b>	<b>(37.9%)</b>
8	107	41 (16.3%)	65	49 (19.4%)	(37.2%)
	<b>107</b>	<b>12 (42.9%)</b>	<b>65</b>	<b>11 (39.3%)</b>	<b>(37.2%)</b>
9	89	44 (17.5%)	47	58 (23.0%)	(36.8%)
	<b>89</b>	<b>11 (39.3%)</b>	<b>47</b>	<b>12 (42.9%)</b>	<b>(36.8%)</b>
10	99	46 (18.3%)	63	55 (21.8%)	(38.8%)
	<b>99</b>	<b>13 (46.4%)</b>	<b>63</b>	<b>12 (42.9%)</b>	<b>(38.8%)</b>

<i>Train (average):</i>	103.2	42.8 (17.0%)	46.0	61.5 (24.4%)	(37.9%)
<i>Test (average):</i>	103.2	13.2 (47.1%)	46.0	13.8 (49.3%)	(37.9%)

Table 6.15: Number of errors and tree sizes for the 2005-2008 dataset: subset 2 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.16 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 16 test cases as UP (second row, last column) and 13 of these were correctly classified (second row, second column). C4.5 classified 12 cases as DOWN (second row, last column) but only 1 of these was correctly classified (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>13</b>	3	(a): class UP	16
	11	<b>1</b>	(b): class DOWN	12
<i>2</i>	<b>8</b>	8	(a): class UP	16
	8	<b>4</b>	(b): class DOWN	12
<i>3</i>	<b>11</b>	5	(a): class UP	16
	12	<b>0</b>	(b): class DOWN	12
<i>4</i>	<b>12</b>	4	(a): class UP	16
	11	<b>1</b>	(b): class DOWN	12
<i>5</i>	<b>12</b>	4	(a): class UP	16
	8	<b>4</b>	(b): class DOWN	12
<i>6</i>	<b>8</b>	8	(a): class UP	16
	8	<b>4</b>	(b): class DOWN	12
<i>7</i>	<b>11</b>	5	(a): class UP	16
	8	<b>4</b>	(b): class DOWN	12
<i>8</i>	<b>10</b>	7	(a): class UP	17
	4	<b>7</b>	(b): class DOWN	11
<i>9</i>	<b>11</b>	6	(a): class UP	17
	6	<b>5</b>	(b): class DOWN	11
<i>10</i>	<b>11</b>	6	(a): class UP	17

	6	5	(b): class DOWN	11
<i>Total</i>	<b>107</b>	56	(a): class UP	163
	82	<b>35</b>	(b): class DOWN	117

Table 6.16: Confusion matrix for the 2005-2008 dataset: subset 2 (C4.5).

If we add all the correct UP classifications, we find that 107/163 UP classifications were correctly classified (65.6%). If we compute a similar value for the DOWNS, we find that 35/117 (29.9%) were correctly classified. A discussion of these results can be found in Section 6.4.5.

### 6.4.3 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (C4.5)

This experiment used the 449 disclosures from the 1997-2000 and 2005-2008 datasets that were also used in the FEP and keyword experiments with C4.5 and SVM-Light (see Sections 6.2.4 and 6.3.4 respectively). The dataset comprised 253 ‘ups’ and 196 ‘downs’, each with one or more automatically extracted five-word n-grams.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 25.3% error rate on the training data and an average 48.1% error rate (or 51.9% accuracy rate) on the unseen test data (see the last two rows in Table 6.17). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	169	67 (16.6%)	109	85 (21.0%)	(39.0%)
	<b>169</b>	<b>21 (46.7%)</b>	<b>109</b>	<b>20 (44.4%)</b>	<b>(39.0%)</b>
2	161	75 (18.6%)	73	106 (26.2%)	(39.5%)
	<b>161</b>	<b>22 (48.9%)</b>	<b>73</b>	<b>23 (51.1%)</b>	<b>(39.5%)</b>
3	161	78 (19.3%)	85	104 (25.7%)	(40.3%)
	<b>161</b>	<b>16 (35.6%)</b>	<b>85</b>	<b>18 (40.0%)</b>	<b>(40.3%)</b>
4	169	78 (19.3%)	63	107 (26.5%)	(38.6%)

	<b>169</b>	<b>24 (53.3%)</b>	<b>63</b>	<b>25 (55.6%)</b>	<b>(38.6%)</b>
5	159	80 (19.8%)	63	112 (27.7%)	(39.7%)
	<b>159</b>	<b>25 (55.6%)</b>	<b>63</b>	<b>26 (57.8%)</b>	<b>(39.7%)</b>
6	171	75 (18.6%)	93	98 (24.3%)	(39.8%)
	<b>171</b>	<b>25 (55.6%)</b>	<b>93</b>	<b>21 (46.7%)</b>	<b>(39.8%)</b>
7	167	76 (18.8%)	71	106 (26.2%)	(39.0%)
	<b>167</b>	<b>21 (46.7%)</b>	<b>71</b>	<b>23 (51.1%)</b>	<b>(39.0%)</b>
8	149	83 (20.5%)	87	100 (24.8%)	(39.8%)
	<b>149</b>	<b>15 (33.3%)</b>	<b>87</b>	<b>19 (42.2%)</b>	<b>(39.8%)</b>
9	147	83 (20.5%)	75	101 (25.0%)	(38.5%)
	<b>147</b>	<b>21 (46.7%)</b>	<b>75</b>	<b>24 (53.3%)</b>	<b>(38.5%)</b>
10	159	85 (21.0%)	89	103 (25.4%)	(40.9%)
	<b>159</b>	<b>19 (43.2%)</b>	<b>89</b>	<b>17 (38.6%)</b>	<b>(40.9%)</b>
<i>Train (average):</i>	161.2	78.0 (19.3%)	80.8	102.2 ( <b>25.3%</b> )	(39.5%)
<i>Test (average):</i>	161.2	20.9 (46.6%)	80.8	21.6 ( <b>48.1%</b> )	(39.5%)

Table 6.17: Number of errors and tree sizes for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.18 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 25 test cases as UP (second row, last column) and 15 of these were correctly classified (second row, second column). C4.5 classified 20 cases as DOWN (second row, last column) and 10 of these were correctly classified (second row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
1	<b>15</b>	10	(a): class UP	25
	10	<b>10</b>	(b): class DOWN	20
2	<b>15</b>	10	(a): class UP	25
	13	<b>7</b>	(b): class DOWN	20
3	<b>11</b>	14	(a): class UP	25
	4	<b>16</b>	(b): class DOWN	20
4	<b>13</b>	12	(a): class UP	25

	13	<b>7</b>	(b): class DOWN	20
5	<b>14</b>	11	(a): class UP	25
	15	<b>5</b>	(b): class DOWN	20
6	<b>15</b>	10	(a): class UP	25
	11	<b>9</b>	(b): class DOWN	20
7	<b>13</b>	12	(a): class UP	25
	11	<b>9</b>	(b): class DOWN	20
8	<b>14</b>	12	(a): class UP	26
	7	<b>12</b>	(b): class DOWN	19
9	<b>14</b>	12	(a): class UP	26
	12	<b>7</b>	(b): class DOWN	19
10	<b>15</b>	11	(a): class UP	26
	6	<b>12</b>	(b): class DOWN	18
<i>Total</i>	<b>139</b>	114	(a): class UP	253
	102	<b>94</b>	(b): class DOWN	196

Table 6.18: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (C4.5).

If we add all the correct UP classifications, we find that 139/253 UP classifications were correctly classified (54.9%). If we compute a similar value for the DOWNS, we find that 94/196 (48.0%) were correctly classified. A discussion of these results can be found in Section 6.4.5.

#### 6.4.4 Classification Results for the 2005-2008 Dataset: Subset 3 (C4.5)

This experiment used the 1,246 disclosures from the 2005-2008 dataset that were used in earlier experiments with C4.5 and SVM-Light (see Sections 6.2.5 and 6.3.5 respectively). The dataset comprised 672 ‘ups’ and 574 ‘downs’, each with one or more automatically extracted five-word n-grams.

Using ten-fold cross validation, the C4.5 simplified decision trees had an average 35.4% error rate on the training data and an average 49.6% error rate (or 50.4% accuracy rate) on the unseen test data (see the last two rows in Table 6.19). The table also shows the number of errors for training and test data, for each fold. For each fold, the rows in bold type signify the evaluations on unseen test cases.

<i>Cross-validation fold</i>	<i>Size before pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Size after pruning (number of nodes)</i>	<i>Number of errors on training or test data</i>	<i>Estimated error for training data</i>
1	237	352 (31.4%)	83	398 (35.5%)	(43.0%)
	<b>237</b>	<b>55 (44.0%)</b>	<b>83</b>	<b>60 (48.0%)</b>	<b>(43.0%)</b>
2	259	360 (32.1%)	53	439 (39.2%)	(44.1%)
	<b>259</b>	<b>53 (42.4%)</b>	<b>53</b>	<b>62 (49.6%)</b>	<b>(44.1%)</b>
3	267	332 (29.6%)	121	375 (33.5%)	(42.7%)
	<b>267</b>	<b>78 (62.4%)</b>	<b>121</b>	<b>69 (55.2%)</b>	<b>(42.7%)</b>
4	257	335 (29.9%)	123	381 (34.0%)	(43.4%)
	<b>257</b>	<b>61 (48.8%)</b>	<b>123</b>	<b>61 (48.8%)</b>	<b>(43.4%)</b>
5	251	336 (30.0%)	85	403 (36.0%)	(43.1%)
	<b>251</b>	<b>67 (53.6%)</b>	<b>85</b>	<b>67 (53.6%)</b>	<b>(43.1%)</b>
6	251	337 (30.1%)	119	384 (34.3%)	(43.5%)
	<b>251</b>	<b>60 (48.0%)</b>	<b>119</b>	<b>63 (50.4%)</b>	<b>(43.5%)</b>
7	251	340 (30.3%)	69	412 (36.7%)	(43.3%)
	<b>251</b>	<b>60 (48.4%)</b>	<b>69</b>	<b>67 (54.0%)</b>	<b>(43.3%)</b>
8	231	351 (31.3%)	99	396 (35.3%)	(43.6%)
	<b>231</b>	<b>51 (41.1%)</b>	<b>99</b>	<b>56 (45.2%)</b>	<b>(43.6%)</b>
9	255	352 (31.4%)	97	395 (35.2%)	(43.3%)
	<b>255</b>	<b>53 (42.7%)</b>	<b>97</b>	<b>50 (40.3%)</b>	<b>(43.3%)</b>
10	249	345 (30.7%)	115	384 (34.2%)	(43.2%)
	<b>249</b>	<b>62 (50.0%)</b>	<b>115</b>	<b>63 (50.8%)</b>	<b>(43.2%)</b>
<i>Train (average):</i>	250.8	344.0 (30.7%)	96.4	396.7 (35.4%)	(43.3%)
<i>Test (average):</i>	250.8	60.0 (48.1%)	96.4	61.8 (49.6%)	(43.3%)

Table 6.19: Number of errors and tree sizes for the 2005-2008 dataset: subset 3 (C4.5).

Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.20 provides a confusion matrix for the simplified trees on the test cases, showing how the misclassifications were distributed. In the first fold, C4.5 classified 67 test cases as UP (second row, last column) and 50 of these were correctly classified (second row, second column). C4.5 classified 58 cases as DOWN (second row, last column) but only 15 of these were correctly classified (second row, third column). In

the second and third columns, figures highlighted in bold indicate correct classifications.

<i>Fold</i>	<i>(a)</i>	<i>(b)</i>	<i>Classified as</i>	<i>Total # classifications (C4.5)</i>
<i>1</i>	<b>50</b>	17	(a): class UP	67
	43	<b>15</b>	(b): class DOWN	58
<i>2</i>	<b>49</b>	18	(a): class UP	67
	44	<b>14</b>	(b): class DOWN	58
<i>3</i>	<b>40</b>	27	(a): class UP	67
	42	<b>16</b>	(b): class DOWN	58
<i>4</i>	<b>47</b>	20	(a): class UP	67
	41	<b>17</b>	(b): class DOWN	58
<i>5</i>	<b>41</b>	26	(a): class UP	67
	41	<b>17</b>	(b): class DOWN	58
<i>6</i>	<b>40</b>	27	(a): class UP	67
	36	<b>22</b>	(b): class DOWN	58
<i>7</i>	<b>37</b>	30	(a): class UP	67
	37	<b>20</b>	(b): class DOWN	57
<i>8</i>	<b>47</b>	20	(a): class UP	67
	36	<b>21</b>	(b): class DOWN	57
<i>9</i>	<b>49</b>	19	(a): class UP	68
	31	<b>25</b>	(b): class DOWN	56
<i>10</i>	<b>44</b>	24	(a): class UP	68
	39	<b>17</b>	(b): class DOWN	56
<i>Total</i>	<b>444</b>	228	(a): class UP	672
	390	<b>184</b>	(b): class DOWN	574

Table 6.20: Confusion matrix for the 2005-2008 dataset: subset 3 (C4.5).

If we add all the correct UP classifications, we find that 444/672 UP classifications were correctly classified (66.1%). If we compute a similar value for the DOWNS, we find that 184/574 (32.1%) were correctly classified. A discussion of these results can be found in the next section.

#### 6.4.5 Analysis and Discussion: C4.5 Results

Looking at the error rates on training data for all four n-gram experiments, it ranged from 17.9% (1997-2000 dataset, subset 1) to 35.4% (2005-2008 dataset, subset 3), averaging at about 25.8%. In the summary to this chapter (Section 6.7), we will compare these relatively high error rates with those from the equivalent C4.5 experiments where we used FEPs and keywords rather than n-grams.

In the 1997-2000 dataset (Section 6.4.1), the average accuracy on ‘up’ and ‘down’ test cases was 53.3%. In the 2005-2008 experiment (Section 6.4.2), the average accuracy was 50.7%. If we were to randomly assign each disclosure to the ‘up’ or ‘down’ category, we might expect a 50-50 result; in this case, use of n-grams and C4.5 would outperform chance for both experiments, although the figure for the 2005-2008 experiment is only marginally better-than-chance.

On closer examination of the ‘ups’ and ‘downs’, we found that the classifiers correctly classified 47.8% of the ‘ups’ and 59.5% of the ‘downs’ (1997-2000 experiment, see Section 6.4.1). As discussed earlier, a more appropriate baseline approach is to compare these figures with arbitrary classification, taking the actual proportion of ‘ups’ and ‘downs’ in our datasets into account. In the 1997-2000 experiment, 53.3% of the dataset comprised ‘ups’ so n-gram classification did not yield improved results over arbitrary classification (compare 47.8% with 53.3%). As regards the ‘downs’, which comprised 46.7% of the dataset, n-gram classification did yield improved results over arbitrary classification (compare 59.5% with 46.7%).

If we conduct a similar analysis on the 2005-2008 dataset, where the classifiers correctly classified 65.6% of the ‘ups’ and 29.9% of the ‘downs’, we find that n-gram classification outperformed arbitrary classification with the ‘ups’ (compare 65.6% accuracy with the actual proportion of ‘ups’ which was 58.2%) but it did not outperform arbitrary classification with the ‘downs’ (compare 29.9% accuracy with the actual proportion of ‘downs’ which was 41.8%).

In Section 6.4.3, we discussed the accuracy of n-gram classification when we merged the 1997-2000 and 2005-2008 datasets. We found that 54.9% of the ‘ups’ were

correctly classified. N-gram classification did not outperform arbitrary classification because the actual proportion of ‘ups’ was 56.3%. Regarding the ‘downs’, n-gram classification did outperform arbitrary classification as the accuracy was 48.0% which compares favourably with the actual proportion of ‘downs’ which was 43.7%.

Finally, in Section 6.4.4, we discussed the results for 1,246 8-Ks filed for 50 S & P companies during 2005-2008. The average accuracy for ‘ups’ was 66.1%, which compares favourably with the results that would be achieved using arbitrary classification (53.9%). Regarding the ‘downs’, the average accuracy was only 32.1% which falls well below the result that would be achieved with arbitrary classification (46.1%).

To summarise, if we compare the classification accuracy of ‘ups’ and ‘downs’ separately with the actual proportion of ‘ups’ and ‘downs’ in each experiment, the results for C4.5 using n-grams are as follows:

*1997-2000 dataset (subset 1):*

- C4.5: ‘ups’ does not outperform arbitrary classification.
- C4.5: ‘downs’ **outperforms** arbitrary classification.

*2005-2008 dataset (subset 2):*

- C4.5: ‘ups’ **outperforms** arbitrary classification.
- C4.5: ‘downs’ does not outperform arbitrary classification.

*1997-2000 and 2005-2008 datasets (subsets 1 and 2):*

- C4.5: ‘ups’ does not outperform arbitrary classification.
- C4.5: ‘downs’ **outperforms** arbitrary classification.

*2005-2008 dataset (subset 3):*

- C4.5: ‘ups’ **outperforms** arbitrary classification.
- C4.5: ‘downs’ does not outperform arbitrary classification.

Summarising these results, we can say that C4.5 outperforms arbitrary classification with the ‘ups’ in two of the four experiments, when n-grams were used as document features. In both cases, the data was from the 2005-2008 period, which could suggest that n-grams were more useful predictive features for ‘ups’ after the 2004 rule changes. As regards the ‘downs’, C4.5 also outperforms arbitrary classification on two occasions—once with the 1997-2000 dataset but also when pre- and post-2004 datasets were combined. Therefore, we cannot definitely say that n-grams are more or less useful predictive features for ‘downs’ pre- or post- the rule changes.

## **6.5 Support Vector Machine Classification using N-Grams**

In Section 6.3.1 we presented an overview of classification using support vector machines, with a specific focus on the SVM-Light program. In Section 6.4, we outlined how we automatically extracted the n-grams for our experiments.

In Sections 6.5.1 to 6.5.4, we present the results from a number of n-gram experiments which used SVM-Light for classification:

- Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light)
- Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light)
- Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM-Light)
- Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light)

In Section 6.5.5, we will analyse and discuss the results from each of the four experiments.

### **6.5.1 Classification Results for the 1997-2000 Dataset: Subset 1 (SVM-Light)**

This experiment used the 169 disclosures from the 1997-2000 dataset that were also used in the experiments described in Sections 6.2.2 (C4.5 with FEPs and keywords), 6.3.2 (SVM-Light with FEPs and keywords), and 6.4.1 (C4.5 with n-grams). The dataset comprised 90 ‘ups’ and 79 ‘downs’, each comprising one or more automatically extracted five-word n-grams.

Using our method of ten-fold cross validation, SVM-Light had an average 28.1% error rate on the training data and an average 46.8% error rate (or 53.2% accuracy rate) on the unseen test data (see the last row and column in Table 6.21). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 21.1% of the ‘ups’ (78.9% accuracy) but 72.5% of the ‘downs’ (27.5% accuracy).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	135	41 (30.4%)	$\leq 91.85\%$	18 (u) 16 (d)	2 (11.1%) 12 (75.0%)
2	135	41 (30.4%)	$\leq 94.07\%$	18 (u) 16 (d)	3 (16.7%) 11 (68.8%)
3	135	36 (26.7%)	$\leq 94.81\%$	18 (u) 16 (d)	4 (22.2%) 12 (75.0%)
4	135	38 (28.1%)	$\leq 94.81\%$	18 (u) 16 (d)	3 (16.7%) 12 (75.0%)
5	135	42 (31.1%)	$\leq 90.37\%$	18 (u) 16 (d)	4 (22.2%) 13 (81.3%)
6	135	29 (21.5%)	$\leq 91.11\%$	18 (u) 16 (d)	5 (27.8%) 12 (75.0%)
7	135	32 (23.7%)	$\leq 94.07\%$	18 (u) 16 (d)	4 (22.2%) 9 (56.3%)
8	136	42 (30.9%)	$\leq 94.12\%$	18 (u) 15 (d)	4 (22.2%) 11 (73.3%)
9	135	38 (28.1%)	$\leq 94.07\%$	18 (u) 16 (d)	6 (33.3%) 11 (68.8%)
10	135	39 (28.9%)	$\leq 95.56\%$	18 (u) 16 (d)	3 (16.7%) 13 (81.3%)
<i>Average</i>	135	38 (28.1%)	$\leq 93.48\%$	18 (u) 16 (d)	3.8 (21.1%) 11.6 (72.5%)

Table 6.21: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 dataset: subset 1 (SVM-Light).

### 6.5.2 Classification Results for the 2005-2008 Dataset: Subset 2 (SVM-Light)

This experiment used the 280 disclosures that were also used in the experiments described in Sections 6.2.3 (C4.5 with FEPs and keywords), 6.3.3 (SVM-Light with FEPs and keywords), and 6.4.2 (C4.5 with n-grams). The dataset comprised 163 ‘ups’ and 117 ‘downs’, each comprising one or more automatically extracted five-word n-grams.

Using our method of ten-fold cross validation, SVM-Light had an average 39.3% error rate on the training data and an average 49.0% error rate (or 51.0% accuracy rate) on the unseen test data (see the last row and column in Table 6.22). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 5.5% of the ‘ups’ (94.5% accuracy) but 92.5% of the ‘downs’ (7.5% accuracy).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	224	85 (37.9%)	<=82.59%	33 (u) 23 (d)	5 (15.2%) 21 (91.3%)
2	224	93 (41.5%)	<=81.70%	32 (u) 24 (u)	0 (0%) 24 (100%)
3	224	88 (39.3%)	<=82.59%	32 (u) 24 (d)	3 (9.4%) 24 (100%)
4	223	93 (41.7%)	<=82.06%	33 (u) 24 (d)	0 (0%) 24 (100%)
5	224	86 (38.4%)	<=83.48%	33 (u) 23 (d)	3 (9.1%) 21 (91/3%)
6	225	90 (40.0%)	<=82.22%	32 (u) 23 (d)	1 (3.1%) 23 (100%)
7	224	88 (39.3%)	<=82.59%	33 (u) 23 (d)	3 (9.1%) 23 (100%)
8	223	82 (36.8%)	<=81.61%	33 (u) 24 (d)	1 (3.0%) 19 (79.2%)
9	224	86 (38.4%)	<=82.14%	33 (u) 24 (d)	1 (3.0%) 19 (79.2%)
10	223	91 (40.8%)	<=82.06%	33 (u) 24 (d)	1 (3.0%) 24 (100%)
<i>Average</i>	224	88 (39.3%)	<=82.30%	33 (u) 24 (d)	1.8 (5.5%) 22.2 (92.5%)

Table 6.22: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 2 (SVM-Light).

### 6.5.3 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (SVM-Light)

This experiment used the 449 disclosures that were also used in the experiments described in Sections 6.2.4 (C4.5 with FEPs and keywords), 6.3.4 (SVM-Light with FEPs and keywords), and 6.4.3 (C4.5 with n-grams). The data comprised 253 ‘ups’ and 196 ‘downs’, each comprising one or more automatically extracted five-word n-grams.

Using our method of ten-fold cross validation, SVM-Light had an average 38.7% error rate on the training data and an average 50.0% error rate (or 50.0% accuracy rate) on the unseen test data (see the last row and column in Table 6.23). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 8.6% of the ‘ups’ (91.4% accuracy) but 91.3% of the ‘downs’ (8.7% accuracy).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	360	136 (37.8%)	$\leq 87.22\%$	50 (u) 39 (d)	9 (18.0%) 34 (87.2%)
2	358	140 (39.1%)	$\leq 86.87\%$	51 (u) 40 (d)	1 (2.0%) 38 (95.0%)
3	359	133 (37.0%)	$\leq 86.91\%$	51 (u) 39 (d)	6 (11.8%) 34 (87.2%)
4	359	129 (35.9%)	$\leq 86.91\%$	51 (u) 39 (d)	5 (9.8%) 34 (87.2%)
5	358	146 (40.8%)	$\leq 86.87\%$	51 (u) 40 (d)	3 (5.9%) 40 (100%)
6	359	135 (37.6%)	$\leq 87.47\%$	51 (u) 39 (d)	12 (23.5%) 33 (84.6%)
7	359	150 (41.8%)	$\leq 87.47\%$	51 (u) 39 (d)	1 (2.0%) 39 (100%)
8	358	131 (36.6%)	$\leq 86.59\%$	51 (u) 40 (d)	1 (2.0%) 34 (85.0%)
9	358	140 (39.1%)	$\leq 86.87\%$	51 (u) 40 (d)	1 (2.0%) 34 (85.0%)
10	359	146 (40.7%)	$\leq 87.19\%$	51 (u) 39 (d)	5 (9.8%) 36 (92.3%)
<i>Average</i>	359	139 (38.7%)	$\leq 87.04$	51 (u) 39 (d)	4.4 (8.6%) 35.6 (91.3%)

Table 6.23: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (SVM-Light).

#### 6.5.4 Classification Results for the 2005-2008 Dataset: Subset 3 (SVM-Light)

This experiment used the 1,246 disclosures that were filed for 50 S&P 500 companies during the period 2005-2008. These disclosures were also used in the experiments described in Sections 6.2.5 (C4.5 with FEPs and keywords), 6.3.5 (SVM-Light with FEPs and keywords), and 6.4.4 (C4.5 with n-grams). The data comprised 672 ‘ups’ and 574 ‘downs’, each comprising one or more automatically extracted five-word n-grams.

Using our method of ten-fold cross validation, SVM-Light had an average 42.9% error rate on the training data and an average 48.5% error rate (or 51.5% accuracy rate) on the unseen test data (see the last row and column in Table 6.24). The table also shows the number of errors for training and test data, for each partition. Looking more closely at the ‘ups’ and ‘downs’, we can see that SVM-Light incorrectly classified only 18.1% of the ‘ups’ (81.9% accuracy) but 78.9% of the ‘downs’ (21.1% accuracy).

<i>Partition</i>	<i># Training cases</i>	<i># Training cases misclassified</i>	<i>XiAlpha-estimate of the error</i>	<i># Up (u) and down (d) test cases</i>	<i># Incorrect up and down classifications</i>
1	997	417 (41.8%)	$\leq 91.07\%$	134 (u) 115 (d)	35 (26.1%) 85 (73.9%)
2	997	417 (41.8%)	$\leq 90.17\%$	134 (u) 115 (d)	39 (29.1%) 86 (74.8%)
3	997	440 (44.1%)	$\leq 90.47\%$	134 (u) 115 (d)	19 (14.2%) 103 (89.6%)
4	997	433 (43.4%)	$\leq 91.07\%$	134 (u) 115 (d)	16 (11.9%) 94 (81.7%)
5	997	429 (43.0%)	$\leq 90.97\%$	134 (u) 115 (d)	27 (20.1%) 87 (75.7%)
6	997	415 (41.6%)	$\leq 90.87\%$	134 (u) 115 (d)	25 (18.7%) 94 (81.7%)
7	997	425 (42.6%)	$\leq 91.47\%$	134 (u) 115 (d)	25 (18.7%) 87 (75.7%)
8	996	439 (44.0%)	$\leq 91.47\%$	135 (u) 115 (d)	25 (18.5%) 80 (69.6%)
9	998	419 (42.0%)	$\leq 90.88\%$	135 (u) 115 (d)	25 (18.5%) 80 (69.6%)
10	997	442 (44.3%)	$\leq 91.37\%$	134 (u) 115 (d)	7 (5.2%) 111 (96.5%)
<i>Average</i>	997	428 (42.9%)	$\leq 90.98$	134 (u) 115 (d)	24.3 (18.1%) 90.7 (78.9%)

Table 6.24: Number of errors for training cases and ‘up’ and ‘down’ test cases for the 2005-2008 dataset: subset 3 (SVM-Light).

### 6.5.5 Analysis and Discussion: SVM-Light Results

Looking at the error rates on training data for all four experiments, it ranged from 28.1% (1997-2000, subset 1) to 42.9% (2005-2008, subset 3), averaging at 37.3%. Ideally, these error rates need to be significantly lower, if we want to secure low error rates on test data.

In the 1997-2000 experiment (Section 6.5.1), the average accuracy on ‘ups’ and ‘downs’ test cases was 53.2%. In the 2005-2008 experiment (Section 6.5.2), the average accuracy was 51.0%. Comparing these figures to 50-50 chance, we can see that n-grams and SVM-Light yielded slightly better-than-chance results, for both datasets.

For the first experiment, if we compare the classification accuracy of ‘ups’ and ‘downs’ to a arbitrary classification, we find that SVM-Light and n-grams perform better for the ‘ups’ (compare 78.9% accuracy with 53.3% arbitrary classification) but not for the ‘downs’ (compare 27.5% accuracy with 46.7% arbitrary classification). In the second experiment, we find that SVM-Light and n-grams also perform better for the ‘ups’ (compare 94.5% accuracy with 58.2% arbitrary classification) but not for the ‘downs’ (compare 7.5% accuracy with 41.8% arbitrary classification).

This pattern continues for the remaining two experiments also. See Table 6.25 the comparison data for each of the four experiments.

	<i>‘ups’ accuracy using n-grams and SVM-Light</i>	<i>‘ups’ accuracy using arbitrary classification</i>	<i>‘downs’ accuracy using n-grams and SVM-Light</i>	<i>‘downs’ accuracy using arbitrary classification</i>
<i>Experiment 1</i>	<b>78.9%</b>	53.3%	27.5%	<b>46.7%</b>
<i>Experiment 2</i>	<b>94.5%</b>	58.2%	7.5%	<b>41.8%</b>
<i>Experiment 3</i>	<b>91.4%</b>	56.3%	8.7%	<b>43.7%</b>
<i>Experiment 4</i>	<b>81.9%</b>	53.9%	21.1%	<b>46.1%</b>

Table 6.25: Accuracy levels of SVM-Light and n-grams compared to arbitrary classification (all experiments).

To summarise, if we compare the actual classification accuracy for ‘ups’ and ‘downs’ separately, with the actual proportion of ‘ups’ and ‘downs’ in each experiment (arbitrary classification), the results for SVM-Light and n-grams are as follows:

*1997-2000 dataset (subset 1):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*2005-2008 dataset (subset 2):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*1997-2000 and 2005-2008 datasets (subsets 1 and 2):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

*2005-2008 dataset (subset 3):*

- SVM-Light: ‘ups’ **outperforms** arbitrary classification.
- SVM-Light: ‘downs’ does not outperform arbitrary classification.

Summarising these results, we can say that SVM-Light always outperforms arbitrary classification with the ‘ups’ but it never outperforms arbitrary classification with the ‘downs’. As discussed in Section 6.3.6 (analysis and discussion of SVM-Light results using FEPs and keywords), SVM-Light behaved like a default classifier, using the largest class (‘up’) as the default class.

## 6.6 Naive Bayes Classification using a Bag-of-Words Approach

In this section, we first present an overview of Naïve Bayes, one of the methods we adopted as a baseline approach. We then present the results from a number of Naïve Bayes experiments in Sections 6.6.2 to 6.6.5:

- Classification Results for the 1997-2000 Dataset: Subset 1 (Naïve Bayes).
- Classification Results for the 2005-2008 Dataset: Subset 2 (Naïve Bayes).
- Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (Naïve Bayes).
- Classification Results for the 2005-2008 Dataset: Subset 3 (Naïve Bayes).

In Section 6.6.6, we will analyse and discuss the results for each of the four experiments.

### 6.6.1 Background to Naïve Bayes

In Section 3.2, we provided an overview of automatic text classification. We also outlined the typical steps required to generate a document representative from an input text, so it can be used by an automatic retrieval system. A document representative can contain features such as single words, n-grams, phrases, terms, or named entities. We also highlighted commonly-used feature selection methods including the bag-of-words approach, information gain, TF\*IDF, and the vector space model. In our experiments, we used a Naïve Bayes (NB) classifier with the bag-of-words representation, as it is a commonly-used baseline approach in the financial domain (Schumaker and Chen 2006).

The Naïve Bayes (NB) classifier assumes that all document attributes (features) are independent of one another given the context of the class; in other words, word order is not relevant (hence the ‘bag-of-words’). Whilst we might not agree with this assumption in the real-world, this assumption has been found to work surprisingly well (McAllum and Nigam 1998). There are several versions of NB models, including the binary independence model (also known as the multi-variate Bernoulli model) and the multinomial model (Lewis 1998; Schneider 2005). The binary model

uses binary word occurrences (the word is either present or absent), whereas the multinomial model uses word occurrence frequencies. We decided to adopt the multinomial model, because it has been found to be more accurate than the binary model, particularly for larger vocabulary sizes, such as those typically found on the Web (McAllum and Nigam 1998) and it accounts for document length within the model (Lewis 1998). Also, the binary model ignores important information that may be inherent in word frequencies (Lewis 1998). We used the Naïve Bayes multinomial classifier provided in the WEKA suite of algorithms<sup>42</sup>.

After removing stopwords and words that occurred less than twice in a document, we calculated word frequencies for every remaining word in every document. We decided not to stem the words on the assumption that stemming may cloud the impact of various words that had the same stems. For example, we may want the classifier to recognise that references to ‘accounts’ may have a negative connotation whereas references to ‘accountant’ may have a positive connotation (or vice-versa).

In Sections 6.6.2 to 6.6.5, we will present the results from four experiments and in Section 6.6.6 we will analyse and discuss the results.

### **6.6.2 Classification Results for the 1997-2000 Dataset: Subset 1 (Naïve Bayes)**

This experiment used the 169 disclosures from the 1997-2000 dataset and comprised 90 ‘ups’ and 79 ‘downs’. Using the bag-of-words approach, each document (case) was represented as a vector of weights, where the weights corresponded to the frequencies of 6,169 words in the various cases.

Using ten-fold cross validation, the Naïve Bayes (NB) algorithm incorrectly classified 47.3% of the cases, yielding a 52.7% accuracy rate. Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.26 provides a confusion matrix, showing how the classifications and misclassifications were distributed. NB correctly classified 35 of the 90 UP cases as UP (second row, second column) and 45 of the 79 DOWN cases as DOWN (third row, third column). In the second and third

---

<sup>42</sup> <http://www.cs.waikato.ac.nz/ml/weka/>

columns, figures highlighted in bold indicate correct classifications.

	(a)	(b)	Classified as	Total # classifications (NB)
<i>Total</i>	<b>35</b>	55	(a): class UP	90
	34	<b>45</b>	(b): class DOWN	79

Table 6.26: Confusion matrix for the 1997-2000 dataset: subset 1 (Naïve Bayes)

### 6.6.3 Classification Results for the 2005-2008 Dataset: Subset 2 (Naïve Bayes)

This experiment used the 280 disclosures from the 2005-2008 dataset, comprising 163 ‘ups’ and 117 ‘downs’. Once again, each case was represented as a vector of weights, where the weights corresponded to the frequencies of 9,218 words in the various cases.

Using ten-fold cross validation, the NB algorithm incorrectly classified 48.9% of the cases, yielding a 51.1% accuracy rate. Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.27 provides a confusion matrix, showing how the classifications and misclassifications were distributed. NB correctly classified 96 of the 163 UP cases as UP (second row, second column) but only 47 of the 117 DOWN cases as DOWN (third row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

	(a)	(b)	Classified as	Total # classifications (NB)
<i>Total</i>	<b>96</b>	67	(a): class UP	163
	70	<b>47</b>	(b): class DOWN	117

Table 6.27: Confusion matrix for the 2005-2008 dataset: subset 2 (Naïve Bayes)

#### 6.6.4 Classification Results for the 1997-2000 and 2005-2008 Datasets: Subsets 1 and 2 (Naïve Bayes)

This experiment used all the data from the previous two experiments (Sections 6.6.2 and 6.6.3). There were 449 disclosures in total, comprising 253 ‘ups’ and 196 ‘downs’. Each case was represented as a vector of weights, where the weights corresponded to the frequencies of 11,803 words in the various cases.

Using ten-fold cross validation, the NB algorithm incorrectly classified 49.4% of the cases, yielding a 50.6% accuracy rate. Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.28 provides a confusion matrix, showing how the classifications and misclassifications were distributed. NB correctly classified 137 of the 253 UP cases as UP (second row, second column) but only 90 of the 196 DOWN cases as DOWN (third row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

	(a)	(b)	Classified as	Total # classifications (NB)
<i>Total</i>	<b>137</b>	116	(a): class UP	253
	106	<b>90</b>	(b): class DOWN	196

Table 6.28: Confusion matrix for the 1997-2000 and 2005-2008 datasets: subsets 1 and 2 (Naïve Bayes).

#### 6.6.5 Classification Results for the 2005-2008 Dataset: Subset 3 (Naïve Bayes)

This experiment used the 1,246 disclosures that were filed for the 50 S & P500 companies, during the period 2005-2008. Only the disclosures (cases) that were <50kb in size were used. There were 672 ‘ups’ and 574 ‘downs’. Each case was represented as a vector of weights, where the weights corresponded to the frequencies of 20,224 words in the various cases.

Using ten-fold cross validation, the NB algorithm incorrectly classified 51.0% of the cases, yielding a 49.0% accuracy rate. Looking more closely at the performance of the ‘ups’ and ‘downs’ separately, Table 6.29 provides a confusion matrix, showing how the classifications and misclassifications were distributed. NB correctly

classified 347 of the 672 UP cases as UP (second row, second column) but only 263 of the 574 DOWN cases as DOWN (third row, third column). In the second and third columns, figures highlighted in bold indicate correct classifications.

	(a)	(b)	Classified as	Total # classifications (NB)
<i>Total</i>	<b>347</b>	325	(a): class UP	672
	311	<b>263</b>	(b): class DOWN	574

Table 6.29: Confusion matrix for the 2005-2008 dataset: subset 3 (Naïve Bayes).

### 6.6.6 Analysis and Discussion: Naïve Bayes Results

If we examine the overall accuracy rates of ups and downs together, we can summarise the results as shown in Table 6.30. Using the bag-of-words approach, the overall accuracy decreased as the datasets increased in size.

	Accuracy
<i>Experiment 1</i>	52.7%
<i>Experiment 2</i>	51.1%
<i>Experiment 3</i>	50.6%
<i>Experiment 4</i>	49.0%

Table 6.30: Overall accuracy of ups and downs together, for all four experiments (Naïve Bayes).

However, if we examine the accuracy of ups and downs separately, the results can be summarised as shown in Table 6.31.

	Ups accuracy	Downs accuracy
<i>Experiment 1</i>	35/90 ( <b>38.9%</b> )	45/79 ( <b>57.0%</b> )
<i>Experiment 2</i>	96/163 ( <b>58.9%</b> )	47/117 ( <b>40.2%</b> )
<i>Experiment 3</i>	137/253 (54.2%)	90/196 (45.9%)
<i>Experiment 4</i>	347/672 (51.6%)	263/574 (45.8%)

Table 6.31: Overall accuracy of ups and downs separately, for all four experiments (Naïve Bayes).

The accuracy of the ‘ups’ was highest in experiment 2 (58.9%) and increased significantly compared to experiment 1 (38.9%), which suggests that the bag-of-words in the ‘ups’ were more informative after the 2004 SEC rule changes. The accuracy decreased from experiments 2 to 4, as the dataset sizes increased. The accuracy of the ‘downs’ was greatest in experiment 1 (57.0%) but decreased significantly in experiment 2 (40.2%). From experiments 2 to 3, the accuracy increased as the dataset size increased, although in experiment 4 there was a marginal decrease in performance (compare 45.8% with 45.9%). A discussion of how these NB results compare to our earlier decision tree and support vector machine experiments can be found in the next section.

## 6.7 Summary

In this chapter, we presented an overview of C4.5 and SVM-Light, the two systems we employed for classification when using FEPs and keywords. We then presented the results of four comparative experiments for each system. We also analysed and discussed the results from each set of experiments. We found that in three of the four experiments, C4.5 equalled or outperformed arbitrary classification with the ‘ups’ and in three of the four experiments, it outperformed arbitrary classification with the ‘downs’. We found that SVM-Light always outperformed arbitrary classification with the ‘ups’ but it never outperformed arbitrary classification with the ‘downs’, possibly because it assigned the larger class (‘up’) as the default. Overall, C4.5 appeared to be better than SVM-Light at identifying patterns in the cases, whether they were ‘up’ or ‘down’ cases, when FEPs and keywords were used. Our findings show that it is possible to achieve results greater-than-chance and arbitrary classification, when using FEPs and keywords as document features for classification.

We also compared the performance of FEP and keyword classification with several baseline approaches: C4.5 classification with n-grams (Section 6.4), SVM-Light classification with n-grams (Section 6.5), and the Naïve Bayes bag-of-words approach to classification (Section 6.6). Our main findings can be summarised as follows:

*Comparing FEP and keyword classification with n-gram classification:*

We found that the n-gram training error for all four C4.5 experiments averaged at 25.8% (see Section 6.4.6). The error rates for these experiments were significantly higher than the equivalent training error rates when using C4.5 with FEPs and keywords (in the latter experiments, the average was 5.1%, see Section 6.2.6). Whilst these training error rates are relatively high, the error rates on *unseen* data were not much greater in the n-gram experiments where accuracy averaged at 51.6% (see Sections 6.4.1 to 6.4.4) compared to the FEP and keywords experiments where accuracy averaged at 52.4% (see Sections 6.2.2. to 6.2.5). In these experiments, the use of FEPs and keywords with C4.5 proved more successful than the use of n-grams, although the difference in accuracy levels was minimal. See Table 6.32 for a summary of the average accuracy levels.

We found that the n-gram training error for all four SVM-Light experiments averaged at 37.3% (see Section 6.5.6). The error rates for these experiments were higher than the equivalent training error rates when using SVM-Light with FEPs and keywords (in the latter experiments the average was 30.0%, see Section 6.3.6). The error rates on unseen data were lower in the n-gram experiments where accuracy averaged at 51.4% (see Sections 6.5.1 to 6.5.4) compared to the FEP and keywords experiments where accuracy averaged at 49.4% (see Sections 6.3.2 to 6.3.5). In these experiments, the use of n-grams with SVM-Light proved more successful than SVM-Light using FEPs and keywords, although we should bear in mind that SVM-Light appeared to use the default classification ('up') on more than one occasion. See Table 6.32 for a summary of these results.

*Comparing FEP and keyword classification with Naïve Bayes (NB) bag-of-words classification:*

Comparing the average classification accuracy on unseen data, we can conclude that the Naïve Bayes bag-of-words approach did not yield improved classification results over C4.5 using FEPs and keywords but it did marginally outperform SVM-Light using FEPs and keywords (see Table 6.32).

<i>C4.5 (FEPs and Keywords) (Section 6.2)</i>	<i>SVM-Light (FEPs and Keywords) (Section 6.3)</i>	<i>C4.5 (N-Grams) (Section 6.4)</i>	<i>SVM-Light (N-Grams (Section 6.5)</i>	<i>Naïve Bayes (Bag-of-Words) (Section 6.6)</i>
<b>52.4%</b>	49.4%	<b>51.6%</b>	51.4%	50.9%

Table 6.32: Average classification accuracy for all four experiments using C4.5, SVM-Light, and Naïve Bayes, with various document content features (FEPs and keywords, n-grams, and a bag-of-words).

So, what can we conclude from all these findings? As shown in Table 6.32, the two highest average accuracy rates were for C4.5 with FEPs and keywords (52.4%) and C4.5 using n-grams (51.6%). Whilst C4.5 using FEPs and keywords (52.4%) performed better than Naïve Bayes bag-of-words classification (50.9%), the relatively small improvement in overall classification accuracy could lead one to argue that this small improvement is not worth the computational effort required to parse disclosures for FEPs, particularly as there is no definitive list of FEPs and recall will suffer as a result.

However, we must also bear in mind that C4.5 found the ‘ups’ easier to classify than the ‘downs’. Consider, for example, the 63.2% classification accuracy achieved for ‘ups’ in the 2005-2008 dataset (subset 2) (see Section 6.2.3). If an investor was to invest solely based on ‘up’ predictions, he/she could potentially make a significant profit. Whilst determining the profitability of various trading strategies is beyond the scope of this research, we do recommend further work in this area, in Chapter 7.

Finally, whilst some aspects of our research are similar to other automatic financial analysis studies that used keywords, n-grams, or the bag-of-words approach (see Chapter 3 for a detailed literature review), these studies are not directly comparable with ours as they either had different goals, used different data sources, and/ or used different methods. Many of these studies also used different evaluation metrics. In the next chapter, we will summarise our findings and draw conclusions from them (Section 7.2). In Section 7.3 we will present the strengths and limitations of our research and in Section 7.4 we will present recommendations for further research.

# Chapter 7: Conclusions and Recommendations

## 7.1 Outline

In the previous chapter, we presented the results from various experiments which used different methods (decision tree classification, support vector machine classification, and Naïve Bayes classification) as well as different content features (FEPs and keywords, n-grams, and a bag-of-words). For each method, we described the systems we chose, before providing the results of our experiments. We then analysed and discussed the results for each method.

In this final chapter, we will summarise our key findings (Section 7.2), outline the strengths and weaknesses of our approach and highlight some important contributions (Section 7.3), and recommend areas worthy of future research (Section 7.4). Finally, in Section 7.5, we will conclude with a summary of the chapter.

## 7.2 Summary of Key Findings

In Chapter 5, we discussed the development of the prototype FEP recogniser. We then undertook a preliminary pattern analysis of two datasets. For both datasets, we examined a number of characteristics, including the total number of words, the number of FEP types recognised, the different FEP types recognised, and the types of FEPs that were duplicated, in ‘downs’ and ‘ups’. During the pattern analysis phase, we found the following:

- There were more than twice as many 8-Ks in the second dataset (2005-2008), compared to the first, but more than three times as many words. We would expect there to be more words in the more recent dataset as companies are now required to file a greater number of event types and therefore it is possible that their 8-Ks will be lengthier.
- In the first dataset (1997-2000), we found that there were proportionately more words in the ‘up’ disclosures<sup>43</sup>, compared to the ‘down’ disclosures (455,730 vs.

---

<sup>43</sup> By ‘up’ disclosures we mean disclosures that had an increase in share price around the filing date.

351,890). In the second dataset, we found that there were proportionately the same number of words in the ‘ups’ and ‘downs’. This suggests that our ‘ups’ contained more verbose language than the ‘downs’ before the 2004 rule changes but that the ‘ups’ and ‘downs’ were fairly similar (word-count wise) after 2004. Either the ‘ups’ became more concise than previously or the ‘downs’ became more verbose.

- Our prototype recogniser only succeeded in identifying one or more FEPs in 30.9% of the ‘downs’ in the first dataset and 20.4% of the ‘downs’ in the second dataset. Similarly, the recogniser only identified FEPs in 30.8% of the ‘ups’ in the first dataset and 24.3% of the ‘ups’ in the second dataset. It seems that the recogniser was less able to identify FEPs in the more recent dataset, even though it had a larger collection of 8-Ks to work with. When examining these figures, we must however bear in mind that these figures relate to FEPs *recognised*—it is quite possible that there were more FEPs in these disclosures. As a result, we recommend further work in this area (see Section 7.4).
- When we examined the different FEP types found by the recogniser, the only FEP type that appeared to be correlated with a *share price increase* is `fep_accountant_appointment`, which only appeared in the ‘ups’ (both datasets).
- When we examined duplicate FEPs, we found that the `fep_new_personnel_or_promotions` was repeated in slightly more ‘downs’ but it was repeated more than twice as often in the second (more recent) dataset. Whilst we would expect to find more occurrences of the FEP in the second dataset, as there are more 8-Ks, we would expect it to appear more in the ‘ups’ than in the ‘downs’. One suggested reason for this might be that this FEP type was repeated more in ‘downs’ to partly disguise other negative news or at least reduce the likely impact. We should point out here that repetition did not affect the classification results however, as FEPs were either present or not present in a disclosure, as far as the classification algorithms were concerned.

In Chapter 6, we presented the results from various experiments which used different methods (decision tree classification, support vector machine classification, and Naïve Bayes classification) as well as different content features (FEPs and keywords, n-grams, and a bag-of-words), for the classification of Form 8-Ks by likely share price response. Here is a brief summary of our results:

- In three of the four experiments, C4.5 equalled or outperformed arbitrary classification with the ‘ups’, when using FEPs and keywords as content features.
- In three of the four experiments, C4.5 outperformed arbitrary classification with the ‘downs’, when using FEPs and keywords as content features.
- SVM-Light always outperformed arbitrary classification with the ‘ups’, when using FEPs and keywords as content features.
- SVM-Light never outperformed arbitrary classification with the ‘downs’, when using FEPs and keywords as content features.
- The C4.5 training error rates with FEPs and keywords were low, ranging from 1.6% to 8.1%.
- The SVM-Light training error rates with FEPs and keywords were high, ranging from 20% to 35.3%.
- In two of the four experiments, C4.5 outperformed arbitrary classification with the ‘ups’, when using n-grams as content features.
- In two of the four experiments, C4.5 outperformed arbitrary classification with the ‘downs’, when using n-grams as content features.
- SVM-Light always outperformed arbitrary classification with the ‘ups’, when using n-grams as content features.
- SVM-Light never outperformed arbitrary classification with the ‘downs’, when using n-grams as content features.
- The C4.5 training error rates with n-grams were quite high, ranging from 17.9% to 35.4%.
- The SVM-Light training error rates with n-grams were quite high, ranging from 28.1% to 42.9%.
- In the Naïve Bayes bag-of-words experiments, the overall classification accuracy of ‘ups’ and ‘downs’ together decreased as the datasets increased in size.
- In the Naïve Bayes bag-of-words experiments, there was no consistent pattern regarding ‘ups’ vs. ‘downs’ classification accuracy, across all four experiments.

So, what do all these results mean from a practical perspective? As shown in table 6.32, the combined average accuracy rates for all experiments only ranged from 49.4% (SVM-Light using FEPs and keywords) to 52.4% (C4.5 using FEPs and keywords), which are not much better than chance. However, if one was to invest solely based on ‘up’ predictions, then C4.5 using FEPs and keywords could potentially yield a profit. As discussed in Section 3.5, Kryzanowski et al (1993) found that negative cases had many more errors than the positive cases when they used the Boltzmann Machine in a two-way classification, but we should point out that there were about five times as many negative cases than positive ones, in their dataset. In our case, we have more ‘ups’ than ‘downs’ but we factored this difference into our evaluation metrics.

As discussed in Section 2.4, Lerman and Livnat (2009) examined the impact of Form 8-K disclosures filed after 2004 to see if their release diminished the impact on other types of disclosures (namely the Form 10-K and Form 10-Q). One of their conclusions was that investors and analysts possibly use these latter disclosure types to interpret 8-Ks filed previously. Another finding was that market reactions varied by event—some event items caused strong positive returns, others caused negative returns. They also suggested that for some events, there may have been an absence of information or inconsistent reactions (events that may have been good news for one firm may have been bad for another). Mittermayer and Knolmayer (2006b) critiqued the level of noise that exists in training data—for example, several Form 8-Ks could be released in a short timeframe, each describing the same event. It would be a reasonable assumption to think that only one of these 8-Ks—presumably the first (and the FEPs contained therein)—actually had an impact on the share price. Classifiers are likely to be impacted by this noisy data. Whilst our evaluation measures show that the ‘up’ disclosures are easier to predict than the ‘down’ disclosures when using C4.5 with FEPs and keywords, it is likely that these same issues impacted our results.

It is clear that more work needs to be done to improve the FEP recognition levels and to identify nuances in negative news disclosures, if we wish to achieve better *overall* classification accuracy for ‘ups’ and ‘downs’. It is possible that certain positive news items typically appear in negative news disclosures to disguise the negative impact,

so it may be possible to identify these. Negative news stories tend to be more difficult to read than positive news stories ( see Section 2.3 ). Another possible explanation could be that there is no correlation between certain event types and share prices.

### 7.3 Strengths and Limitations of our Approach

We believe our research contributes to the field in a number of ways:

- Our study differs from most other studies in that we analyse complete documents and use a combination of attributes ( e.g. financial event phrases, named entities, types of financial object, and keywords) when attempting to classify 8-Ks by likely share price response.
- As outlined in Section 1.2, our prototype FEP recogniser generates output that could be used by investors in their toolbox. We do not claim to have developed the definitive solution; rather, we have identified one possible solution that could facilitate the arduous task of analysing 8-Ks. Investors may be able to make more informed decisions using this output.
- We have identified an extensive list of financial event phrases that could be used for a number of other purposes, including the analysis of online news sources.

In terms of the reliability and validity of our findings, we have considered the following:

- We compared the results from classification experiments which used FEPs and keywords, to the results from other baseline experiments—classification using n-grams (Sections 6.4 and 6.5) and classification using a Naïve Bayes bag-of-words approach (Section 6.6).
- We attempted to avoid sample selection bias by using 50 randomly-chosen companies from the S&P 500 listing, rather than focus on any one industry type.
- The prototype FEP recogniser was developed by the author but ideally investment experts should have been involved, to avoid any possible bias. Nonetheless, we attempted to avoid *event* bias by focusing on all types of financial events rather than specific events only (Antweiler and Frank 2006; Fama 1998). However, the limitations of our prototype recogniser mean that at most 18 of the 49 FEP types

were recognised in disclosures and for some experiments, very few FEP types appeared in the C4.5 decision trees. On the plus side, we did not ignore bad news events just because they may be less frequent (unlike, for example, Liu 2000).

- Fama (1998) warned that many anomalies in reactions tend to vary depending on the models or statistical methodology used to measure them. As outlined in Section 1.3, we have assumed that all companies are equally affected by external factors, so we have not controlled for other variables, such as firm size, book-to-market, institutional ownership, media pessimism, and past trends (Griffin 2003; Tetlock 2007; Loughran et al 2008; Loughran and McDonald 2011b). Also, we did not evaluate measures other than share price return; an alternative measure could be market average return (van Bunningen 2004).
- We believe our findings are robust *to a certain extent*, as we tested several methods using a number of datasets, each comprising four or more years. Also, whilst we only used prices from three-day windows and did not perform sensitivity checks to examine the effects of other windows (unlike, for example, Tetlock (2007), Tetlock et al (2008) and Loughran and McDonald (2011b)), we did consider the wider price dynamics in which 8-Ks were filed (Section 4.4). As outlined in Section 1.3, our study assumes that the market is not fully efficient and prices react within a day to 8-Ks. As discussed in Section 2.4, Hong et al (2000) found that negative firm-specific news in general tends to diffuse more slowly across the investing public, as compared to positive news. Tetlock et al (2008), on the other hand, found that negative news about *earnings* usually causes a reaction within a day. These findings suggest that perhaps some events have quicker (or slower) reaction times than others (Engelberg 2008). A deeper exploration of the effects of other windows might have yielded different results.
- Whilst some argue that intraday prices are better than closing prices (e.g. Fair 2000; Mittermayer and Knolmayer 2006a), it was not possible to determine the specific time of day when the 8-K filings were posted online, so we believed it was best to use a consistent end-of-day pricing method.
- We did not experiment with alternative parameters in C4.5 or SVM-Light, in case our methods became biased by an “inappropriate choice” (Joachims 1998, p.140). However, we could possibly have considered attributes with different costs or used iterative training mode. Whilst we did not test specifically for overfitting,

we did use cross-validation. Also, kernel methods such as support vector machines avoid problems like overfitting, to a certain extent.

- Ideally, with a more efficient FEP recogniser, we would have access to a larger dataset of 8-Ks with one or more recognised FEPs. For automatic classification to have a good chance of success, it is necessary to have sufficient training examples. Use of leave-one-out testing in SVM-Light (Joachims 2000) could prove useful in future when trying to identify the ideal number of training cases.

## 7.4 Further Research

There are a number of avenues that could be taken to build on our research:

- The content of 8-Ks could be examined more deeply, taking into account not just event phrases and keywords, but also themes (e.g. Thomas 1997) or changes in tone (e.g. Feldman et al 2008). These latter characteristics could potentially be used as additional features for classification purposes. A related approach might be to focus on specific event types only, using the findings from previous event studies (see Section 2.4) or the FEPs that appeared in the decision tree output. Also, as discussed in Section 3.3, Loughran and McDonald (2011b) developed a negative word list for financial words—these words could be used to improve our list of keywords, or even to replace it completely. Yet another approach might be to experiment with a combination of FEPs, keywords, and n-grams, to see if the combined features yield better results than they do on their own.
- As outlined in Section 7.3, our prototype FEP recogniser needs further work, firstly to incorporate additional event types (e.g. those added more recently by the SEC, see Table 4.3) so that recall will increase but also to correct any misleading phrases that might affect the precision of the recogniser. Lerman and Livnat (2009) recently identified some events that caused market reactions to 8-Ks, so these could be incorporated into the recogniser also.
- As discussed in Section 2.4, Conrad et al (2002) found that prices tended to respond more strongly to bad news than good news when prices are already high. Kroha et al (2006) said that during a growing trend, optimistic investors may fail to react quickly to negative news; likewise, during a falling trend, pessimistic investors may fail to react quickly to positive news (see Section 3.3 for a

discussion). In Section 4.4, we examined the causal effect of 8-Ks in various windows but as we did not cater for past trends or prices in our experiments, this is also an area that may warrant further research.

- A third classification level could be introduced, such as ‘unclear’ or ‘no recommendation’ (Mittermayer and Kolmayer 2006a). This could reduce the level of noise in the ‘up’ and ‘down’ predictions.
- As proposed by Schumaker and Chen (2006) and implemented by Demers and Vega (2010), another option might be to experiment with more volatile companies or specific industry groups outside the S&P 500.
- Rather than just focus on the directional accuracy (up or down), the potential profitability of the system could be evaluated, taking into account transaction costs, inflation and spread (Malkiel 2007). Fung et al (2005) suggested a buy-and-hold strategy to evaluate the performance. Schumaker and Chen (2006) found that a high directional accuracy does not necessarily imply a highly profitable trading strategy and Qi (1999) said that “predictability does not necessarily imply profitability” (p.425). Whilst our prediction levels equate roughly to chance (if we consider classification of ‘ups’ and ‘downs’ together), it is possible that the better performance of the ‘ups’ could yield some profits. Ultimately, however, one would need to determine if the effort involved in recognising FEPs and classifying 8-Ks, coupled with the associated costs, would be worthwhile in a real-world situation, particularly when n-grams and the bag-of-words yielded broadly comparable results overall (see Section 6.7 for a summary of the results).

## 7.5 Summary

In this chapter, we presented a summary of our key findings. We then outlined the strengths and weaknesses of our approach and highlighted some important contributions of our research. Finally, we recommended a reaseworthy of future research.

## References

- Abarbanell, J. S. and Bernard, V. L. (1992) 'Test of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior', *The Journal of Finance*, 47(3), 1181-1207.
- Ahmad, K., de Oliveira, P. C. F., Manomaisupat, P., Casey, M. and Taskaya Temizel, T. (2002) 'Description of events: A n a nalysis of ke ywords a nd i ndexical names', *Workshop on Event Modelling for Multilingual Document Linking, LREC 2002*, Las Palmas, Canary Islands, Spain, 2nd June 2002.
- Ahmad, K., Gillam, L. and Cheng, D. (2005) 'Textual and quantitative analysis: Towards a new, e-mediated social science', *1st International Conference on e-Social Science*, Manchester, 22-24 June 2005, National Centre for e-Social Science.
- Ahmad, K., Gillam, L., Cheng, D., Taskaya-Temizel, T., Ahmad, S., Manomaisupat, P., Trabloussi, H. and Hippisley, A. (2003) 'The mood of the (financial) markets: In a corpus of words and of pictures', in Archer, D., Rayson, P. R., Wilson, A. and McEnery, T. eds., *Corpus Linguistics 2003 Conference*, Lancaster University, 28-31 March 2003, University Centre for Computer Corpus Research on Language.
- Andersen, T. G. and Bollerslev, T. (1998a) 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review*, 39(4), 885-905.
- Andersen, T. G. and Bollerslev, T. (1998b) 'Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies', *The Journal of Finance*, 53(1), 219-265.
- Antweiler, W. and Frank, M. Z. (2004) 'Is all that talk just noise? The information content of internet stock message boards', *The Journal of Finance*, 59(3), 1259-1294.
- Antweiler, W. and Frank, M. Z. (2006) 'Do U.S. stock markets typically overreact to

corporate news stories?' *Working Paper*, University of British Columbia.

Asthana, S. and Balsam, S. (2001) 'The effect of EDGAR on the market reaction to 10-K filings', *Journal of Accounting and Public Policy*, 20(4-5), 349-372.

Back, B., Toivonen, J., Vanharanta, H. and Visa, A. (2001) 'Comparing numerical data and text information from annual reports using self-organizing maps', *International Journal of Accounting Information Systems*, 2(4), 249-269.

Bagnoli, M., Beneish, M. D. and Watts, S. G. (1999) 'Whisper forecasts of quarterly earnings per share', *Journal of Accounting and Economics*, 28(1), 27-50.

Ball, R. and Brown, P. (1968) 'An empirical evaluation of accounting income numbers', *Journal of Accounting Research*, 6(2), 159-178.

Ball, R. and Kothari, S. P. (1991) 'Security returns around earnings announcements', *The Accounting Review*, 66(4), 718-738.

Banz, R. (1981) 'The relationship between return and market value of common stocks', *Journal of Financial Economics*, 9(1), 3-18.

Barberis, N., Shleifer, A. and Vishny, R. (1998) 'A model of investor sentiment', *Journal of Financial Economics*, 49(3), 307-343.

Barberis, N. and Thaler, R. (2003) 'A survey of behavioral finance' in Constantinides, G. M., Harris, M. and Stulz, R. eds., *Handbook of the economics of finance*: Elsevier Science B.V., 1052-1121.

Basu, S. (1977) 'Investment performance of common stocks in relation to their price-earnings ratio: A test of the efficient market hypothesis', *The Journal of Finance*, 32(3), 663-682.

Beattie, V., McInnes, B. and Fearnley, S. (2004) 'A methodology for analysing and evaluating narratives in annual reports: A comprehensive descriptive profile and metrics for disclosure quality attributes', *Accounting Forum*, 28(3), 205-236.

Bennett, K. P. and Campbell, C. (2000) 'Support vector machines: Hype or

hallelujah?' *ACM SIGKDD Explorations Newsletter - Special Issue on Scalable Data Mining Algorithms*, 2(2), 1-13.

Bernstein, D. (2004) 'Cost of 8-K rules could surpass Sarbanes-Oxley', *International Financial Law Review*, 23(5), 20-24.

Berry, M. J. A. and Linoff, G. (1997) *Data mining techniques: For marketing, sales, and customer support*, New York: John Wiley & Sons.

Berthold, M. and Hand, D. J. (2003) *Intelligent data analysis: An introduction*, 2nd ed., Berlin, Heidelberg: Springer-Verlag.

Brown, S. J. and Warner, J. B. (1985) 'Using daily stock returns: The case of event studies', *Journal of Financial Economics*, 14(1), 3-31.

Brunnermeier, M. K. (1998) *Buy on rumours - sell on news: A manipulative trading strategy*, Discussion Paper 309, London: Financial Markets Group, London School of Economics.

Caropreso, M. F., Matwin, S. and Sebastiani, F. (2001) 'A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization' in Chin, A. G. ed., *Text databases and document management*, Hershey, PA, USA: IGI Publishing.

Carter, M. E. and Soo, B. S. (1999) 'The relevance of Form 8-K reports', *Journal of Accounting Research*, 37(1), 119-132.

Chinchor, N. A. (1998) 'Overview of MUC-7', in Chinchor, N. A. ed., *7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, 29 April - 1 May 1998, San Diego, CA: Science Applications International Corporation.

Cho, V., Wüthrich, B. and Zhang, J. (1999) 'Text processing for classification', *Journal of Computational Intelligence in Finance: Special Issue on Financial News Analysis using Distributed Data Mining*, 7(2), 6-22.

Conrad, J., Cornell, B. and Landsman, W. R. (2002) 'When is bad news really bad news?' *The Journal of Finance*, 57(6), 2507-2532.

- Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(1995), 273-297.
- Cristianini, N. and Shawe-Taylor, J. (2000) *Support vector machines and other kernel-based learning methods*, Cambridge, UK: Cambridge University Press.
- Cutler, D. M., Poterba, D. M. and Summers, L. H. (1989) 'What moves stock prices?' *Journal of Portfolio Management*, 15(3), 4-12.
- Daly, N., Kearney, C. and Ahmad, K. (2009) 'Correlating market movements with consumer confidence and sentiments: A longitudinal study' in Heyer, G. ed., *Text mining services 2009*, Leipzig, Germany: Leipziger Beiträge zur Informatik.
- Daniel, K., Hirshleifer, D. and Subrahmanyam, A. (1998) 'Investor psychology and security market under- and over-reactions', *The Journal of Finance*, 53 (6), 1839-1885.
- Das, S., Martinez-Jerez, A. and Tufano, P. (2005) 'eInformation: A clinical study of investor discussion and sentiment', *Financial Management*, 34(3), 103-137.
- Das, S. R. and Chen, M. Y. (2001) 'Yahoo! For a mazon: Opinion extraction from small talk on the web', *28th European Finance Association Annual Meeting*, Barcelona, Spain, 22-25 August 2001.
- de Oliveira, P. C. F., Ahmad, K. and Gillam, L. (2002) 'A financial news summarisation system based on lexical cohesion', in Gillam, L. ed., *Making Money in the Financial Services Industry, a Workshop at the Terminology and Knowledge Engineering Conference (TKE 2002)*, Nancy, France, 30 August 2002.
- Demers, E. and Vega, C. (2010) *Soft information in earnings announcements: News or noise?*, International Finance Discussion Papers No. 951, Board of Governors of the Federal Reserve System (U.S.).
- Devitt, A. and Ahmad, K. (2007) 'Sentiment polarity identification in financial news: A cohesion-based approach', *45th Annual Meeting of the Association of*

*Computational Linguistics (ACL 2007)*, Prague, Czech Republic, 25-27 June 2007, Association for Computational Linguistics, 984-991.

Ederington, L. H. and Lee, J. H. (1993) 'How markets process information: News releases and volatility', *The Journal of Finance*, 48(4), 1161-1191.

Edwards, W., Phillips, L. D., Hays, W. L. and Goodman, B. C. (1968) 'Probabilistic information processing systems: Design and evaluation', *IEEE Transactions on Systems Science and Cybernetics*, SSC-4(3), 248-265.

Elton, E. J., Gruber, M. J., Brown, S. J. and Goetzmann, W. N. (2003) *Modern portfolio theory and investment analysis*, New York: John Wiley & Sons, Inc.

Engelberg, J. E. (2008) 'Costly information processing: Evidence from earnings announcements', *American Finance Association Annual Meeting 2009*, San Francisco, CA, 3-5 January 2009.

Faerber, E. (2000) *All about stocks: The easy way to get started*, 2nd ed., New York: McGraw-Hill Professional.

Fagan, J. L. (1987) *Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods*, unpublished thesis (Ph.D.), Cornell University.

Fair, R. C. (2000) *Events that shook the market*, Working Paper No. 00-01, New Haven: Yale International Center for Finance, Yale University.

Fair, R. C. (2003) 'Shock effects on stocks, bonds, and exchange rates', *Journal of International Money and Finance*, 22(2003), 307-341.

Fama, E. F. (1965) 'The behavior of stock-market prices', *The Journal of Business*, 38(1), 35-105.

Fama, E. F. (1970) 'Efficient capital markets: A review of theory and empirical work', *The Journal of Finance*, 25(2), 383-417.

Fama, E. F. (1991) 'Efficient capital markets II', *The Journal of Finance*, 46(5), 1575-1617.

- Fama, E. F. (1998) 'Market efficiency, long-term returns, and behavioral finance', *Journal of Financial Economics*, 49(3), 283-306.
- Fama, E. F., Fisher, L., Jensen, M. C. and Roll, R. (1969) 'The adjustment of stock prices to new information', *International Economic Review*, 10(1), 1-21.
- Fama, E. F. and French, K. R. (1992) 'The cross-section of expected stock returns', *The Journal of Finance*, 47(2), 427-465.
- Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2008) 'The incremental information content of tone change in management discussion and analysis', *Working paper*, Available at: <http://ssrn.com/abstract=1126962>.
- Flesch, R. (1972) *The art of readable writing*, New York: Collier.
- Fradkin, D. and Muchnik, I. (2006) 'Support vector machines for classification' in Abello, J. and Cormode, G. eds., *Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 70, 13-20.
- Frakes, W. B. and Baeza-Yates, R., eds. (1992). *Information Retrieval: Data Structures and Algorithms*, New Jersey: Prentice Hall, Englewood Cliffs.
- Francis, J., Schipper, K. and Vincent, L. (2002) 'Earnings announcement and competing information', *Journal of Accounting and Economics*, 33(3), 313-342.
- Fung, G. P. C., Yu, J. X. and Lu, H. (2005) 'The predicting power of textual information on financial markets', *IEEE Intelligent Informatics Bulletin*, 5(1), 1-10.
- Gidófalvi, G. (2001) *Using news articles to predict stock price movements*, Project report, Department of Computer Science and Engineering, University of California, San Diego.
- Gidófalvi, G. and Elkan, C. (2003) *Using news articles to predict stock price movements*, Draft technical report, Department of Computer Science and Engineering, University of California, San Diego.

- Gillam, L., Ahmad, K., Ahmad, S., Casey, M., Cheng, D., Taskaya Temizel, T., de Oliveira, P. C. F. and Manomaisupat, P. (2002) 'Economic news and stock market correlation: A study of the UK market', in Gillam, L. ed., *Making Money in the Financial Services Industry, a Workshop at the Terminology and Knowledge Engineering Conference (TKE 2002)*, Nancy, France, 30 August 2002.
- Goodhart, C. A. E. and O'Hara, M. (1997) 'High frequency data in financial markets: Issues and applications', *Journal of Empirical Finance*, 4, 73-114.
- Griffin, P. A. (2003) 'Got information? Investor response to form 10-K and form 10-Q EDGAR filings', *Review of Accounting Studies*, 8(4), 433-460.
- Gunning, R. (1952) *The technique of clear writing*, New York: McGraw-Hill.
- Halliday, M. A. K. (1985) *An introduction to formal grammar*, London: Edward Arnold.
- Harries, M. and Horn, K. (1995) 'Detecting concept drift in financial time series prediction using symbolic machine learning', *In Proceedings of the Eight Australian Joint Conference on Artificial Intelligence*, Singapore, World Scientific, 91-98.
- Haugen, R. A. (1990) *Modern investment theory*, 2nd ed., New Jersey: Prentice Hall, Englewood Cliffs.
- Hellstrom, T. and Holmstrom, K. (1998) *Predicting the stock market*, Technical Report Series IMA-TOM-1997-07, Mälardalen University, Sweden.
- Henry, E. (2008) 'Are investors influenced by how earnings press releases are written?' *Journal of Business Communication*, 45(4), 363-407.
- Hildebrandt, H. W. and Snyder, R. D. (1981) 'The pollyanna hypothesis in business writing: Initial results, suggestions for research', *The Journal of Business Communication*, 18(1), 5-15.
- Hong, H., Lim, T. and Stein, J. C. (2000) 'Bad news travels slowly: Size, analyst

coverage, and the profitability of momentum strategies', *The Journal of Finance*, 55(1), 265-295.

Ingargiola, G. date unknown 'Building classification models: ID3 and C4.5', *CIS 587: Introduction to Artificial Intelligence*,

Joachims, T. (1998) 'Text categorization with support vector machines: Learning with many relevant features', in Nedellec, C. and Rouveirol, C. eds., *10th European Conference on Machine Learning (ECML-98)*, Chemnitz, Germany, 21 -24 April 1998, Berlin, Heidelberg: Springer-Verlag, 137-142.

Joachims, T. (2000) 'Estimating the generalization performance of a SVM efficiently', in Langley, P. ed., *17th International Conference on Machine Learning (ICML-2000)*, Stanford University, 29 June - 2 July 2000, San Francisco, CA: Morgan Kaufman, 431-438.

Keane (2010) 'EDGAR public dissemination service technical specification', [online], available: <http://www.sec.gov/info/edgar/pdsdissemspec051310.pdf>, [accessed 27th March 2012].

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H. and Visa, A. (2004) 'Combining data and text mining techniques for analyzing financial reports', *Intelligent Systems in Accounting, Finance and Management*, 12(1), 29-41.

Koh, H. C. and Low, C. K. (2004) 'Going concern prediction using data mining techniques', *Managerial Auditing Journal*, 19(3), 462-476.

Kohut, G. F. and Segars, A. H. (1992) 'The president's letter to stockholders: An examination of corporate communication strategy', *The Journal of Business Communication*, 29(1), 7-21.

Koppel, M. and Schrimberg, I. (2004) 'Good news or bad news? Let the market decide', in Qu, Y., Shanahan, J. and Wiebe, J. eds., *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Palo Alto, California, 22-24 March 2004, Menlo Park, CA: The AAAI Press, 86-88.

- Kosala, R. and Blockeel, H. (2000) 'Web mining research: A survey', *ACM SIGKDD Explorations Newsletter*, 2(1), 1-15.
- Kroha, P. and Baeza-Yates, R. (2004) *Classification of stock exchange news*, Technical Report, Department of Computer Science, Engineering School, Universidad de Chile.
- Kroha, P., Baeza-Yates, R. and Krellner, B. (2006) 'Text mining of business news for forecasting', in Bressan, S., Küng, J. and Wagner, R. eds., *17th International Conference on Database and Expert Systems Applications (DEXA '06)*, Krakow, Poland, 4-8 September 2006, Berlin, Heidelberg: Springer-Verlag, 171-175.
- Kryzanowski, L., Galler, M. and Wright, D. W. (1993) 'Using artificial neural networks to pick stocks', *Financial Analysts Journal*, 49(4), 21-27.
- Lam, M. (2004) 'Neural network techniques for financial performance prediction: Integrating fundamental and technical analysis', *Decision Support Systems*, 37(4), 567-581.
- Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J. (2000) 'Language models for financial news recommendation', in Agah, A., Callan, J. and Rundensteiner, E. eds., *Ninth International Conference on Information and Knowledge Management (CIKM)*, McLean, Virginia, USA, 6-11 November 2000, ACM Press, 389-396.
- Lerman, A. and Livnat, J. (2009) 'The new Form 8-K disclosures', *Review of Accounting Studies*, 15(4), 752-778.
- Lewis, D. D. (1992a) 'Feature selection and feature extraction for text categorization', *Workshop on Speech and Natural Language*, Harriman, New York, 23-26 February, Morgan Kaufmann, 212-217.
- Lewis, D. D. (1992b) *Representation and learning in information retrieval*, unpublished thesis (Ph.D.), University of Massachusetts.
- Lewis, D. D. (1998) 'Naive (Bayes) at forty: The independence assumption in

- information retrieval', *Lecture Notes in Computer Science*, 1398/1998, 4-15.
- Li, F. (2006) 'Do stock market investors understand the risk sentiment of corporate annual reports?' *Unpublished working paper*, University of Michigan.
- Li, F. (2008) 'Annual report readability, current earnings, and earnings persistence', *Journal of Accounting and Economics*, 45(2-3), 221-247.
- Liu, A., Y., Gu, B., Konana, P. and Ghosh, J. (2006) *Predicting stock price from financial message boards with a mixture of experts framework*, Technical report, Intelligent Data Exploration & Analysis Laboratory (IDEAL), Department of Electrical and Computer Engineering, The University of Texas at Austin, Texas.
- Liu, Q. (2000) 'How good is good news? Technology depth, book-to-market ratios, and innovative events', *Working Paper*, Department of Economics, University of California, Los Angeles.
- Lo, A. W. and MacKinlay, A. C. (1999) *A non-random walk down wall street*, Princeton: Princeton University Press.
- Loughran, T. and McDonald, B. (2011a) 'Measuring readability in financial disclosures', *Summer Finance Seminar Series*, University of Notre Dame, 17th August 2011.
- Loughran, T. and McDonald, B. (2011b) 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks', *The Journal of Finance*, 66(1), 35-65.
- Loughran, T., McDonald, B. and Yun, H. (2008) 'A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports', *Journal of Business Ethics*, 89(1), 39-49.
- Luhn, H. P. (1958) 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2(2), 159-165.
- MacKinlay, A. C. (1997) 'Event studies in economics and finance', *Journal of Economic Literature*, 35(1), 13-39.

- Malkiel, B. G. (2007) *A random walk down wall street: The time-tested strategy for successful investing*, 9th ed., New York, London: W.W. Norton & Company.
- Manning, C. D., Raghavan, P. and Schütze, H. (2009) *An introduction to information retrieval*: Cambridge University Press.
- McAllum, A. and Nigam, J. (1998) 'A comparison of event models for naive bayes classification', *AAAI-98 Workshop on Learning for Text Categorization*, AAAI Press, 41-48.
- Mitchell, T. M. (1997) *Machine learning*, Boston, Massachusetts: McGraw-Hill.
- Mittermayer, M.-A. and Knolmayer, G. F. (2006a) 'Newscats: A news categorization and trading system', *Sixth International Conference on Data Mining (ICDM '06)*, Hong Kong, China, 18-22 December 2006, 1002-1007.
- Mittermayer, M.-A. and Knolmayer, G. F. (2006b) *Text mining systems for market response to news: A survey*, Working Paper No 184, Institute of Information Systems, University of Bern.
- MUC (1997) 'MUC-7 information extraction task definition', [online], available: [http://www.itl.nist.gov/iaui/894.02/related\\_projects/muc/proceedings/ie\\_task.html](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ie_task.html), [accessed 27th March 2012].
- Newman, M. L., Pennebaker, J. W., Berry, D. S. and Richards, J. M. (2003) 'Lying words: Predicting deception from linguistic styles', *Personality and Social Psychology Bulletin*, 29(5), 665-675.
- Ng, A. and Fu, A. W.-C. (2003) 'Mining frequent episodes for relating financial events and stock trends', in Whang, K.-Y., Jeon, J., Shim, K. and Srivastava, J. eds., *7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining 2003 (PAKDD 2003)*, Seoul, Korea, 30 April - 2 May 2003, Berlin, Heidelberg: Springer-Verlag, 27-39.
- O'Loughlin, B. and O'Brien, F. (2006) *Fundamentals of investment: An Irish perspective*, Dublin: Gill & Macmillan.

- Ou, J. A. (1990) 'The information content of non-earnings accounting numbers as earnings predictors', *Journal of Accounting Research*, 28(1), 144-163.
- Peramunetilleke, D. and Wong, R. K. (2002) 'Currency exchange rate forecasting from news headlines', *Australian Computer Science Communications*, 24(2), 131-139.
- Pereira, F., C.N. and Warren, D. H. D. (1980) 'Definite clause grammars for language analysis--a survey of the formalism and a comparison with augmented transition networks', *Artificial Intelligence*, 13(3), 231-278.
- Pesaran, M. H. and Timmerman, A. (1992) 'A simple nonparametric test of predictive performance', *Journal of Business and Economic Statistics*, 10(4), 461-465.
- Qi, M. (1999) 'Nonlinear predictability of stock returns using financial and economic variables', *Journal of Business and Economic Statistics*, 17(4), 419-429.
- Quinlan, J. R. (1993) *C4.5: Programs for machine learning*, San Mateo, CA: Morgan Kaufmann.
- Racine, J. (2001) 'On the nonlinear predictability of stock returns using financial and economic variables', *Journal of Business and Economic Statistics*, 19(3), 380-382.
- Reinganum, M. R. (1988) 'The anatomy of a stock market winner', *Financial Analysts Journal*, 44(2), 16-28.
- Robles-Granda, P. D. and Belik, I. V. (2010) 'A comparison of machine learning classifiers applied to financial datasets', *In Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, San Francisco, USA, 20 - 22 October 2010.
- S&P (2012) 'S&P 500', [online], available: <http://www.standardandpoors.com/indices/sp-500/en/us/?indexId=spusa-500-usduf--p-us-l->, [accessed 26th March 2012].
- Salton, G. (1970) 'Automatic text analysis', *Science*, 168(3929), 335-343.

- Schneider, K.-M. (2005) 'Techniques for improving the performance of naive bayes for text classification', in Gelbukh, A. ed., *6th International Conference on Intelligent Text Processing and Computational Linguistic (CICLing 2005)*, Mexico City, Mexico, Berlin, Heidelberg: Springer-Verlag, 682-693.
- Schumaker, R. P. and Chen, H. (2006) 'Textual analysis of stock market prediction using financial news articles', in Rodriguez-Abitia, G. and Ania, B., Ignacio eds., *12th Americas Conference on Information Systems (AMCIS-2006)*. Acapulco, Mexico, 4-6 August 2006.
- SEC (1998) *A plain English handbook*, Office of Investor Education and Assistance, U.S. Securities and Exchange Commission [online], available: <http://www.sec.gov/pdf/handbook.pdf>, August 1998.
- SEC (2002) 'Proposed rule: Additional Form 8-K disclosure requirements and acceleration of filing date. Release nos. 33-8106 and 34-46084', [online], available: <http://www.sec.gov/rules/proposed/33-8106.htm>, [accessed 27th March 2012].
- SEC (2004a) 'Final rule: Additional Form 8-K disclosure requirements and acceleration of filing date', [online], available: <http://www.sec.gov/rules/final/33-8400.htm>, [accessed 27th March 2012].
- SEC (2004b) 'Final rule: Asset-backed securities', [online], available: <http://www.sec.gov/rules/final/33-8518.htm>, [accessed 27th March 2012].
- SEC (2005) 'Final rule: Use of Form S-8, Form 8-K, and Form 20-F by shell companies', [online], available: <http://www.sec.gov/rules/final/33-8587.pdf>, [accessed 27th March 2012].
- SEC (2009a) 'Final rule: Proxy disclosure enhancements', [online], available: <http://www.sec.gov/rules/final/2009/33-9089.pdf>, [accessed 27th March 2012].
- SEC (2009b) 'Interactive data to improve financial reporting: Final rule', [online], available: <http://www.sec.gov/rules/final/2009/33-9002fr.pdf>, [accessed 27th March 2012].

- SEC (2010a) 'Form 8-K', [online], available: <http://www.sec.gov/about/forms/form8-k.pdf>, [accessed 27th March 2012].
- SEC (2010b) 'The investor's advocate: How the SEC protects investors, maintains market integrity, and facilitates capital formation', [online], available: <http://www.sec.gov/about/whatwedo.shtml>, [accessed 27th March 2012].
- SEC (2011) 'Adoption of updated EDGAR filing manual', [online], available: <http://www.sec.gov/rules/final/2011/33-9169.pdf>, [accessed 27th March 2012].
- Segars, A. H. and Kohut, G. F. (2001) 'Strategic communication through the world wide web: A new empirical model of effectiveness in the CEO's letter to shareholders', *Journal of Management Studies*, 38(4), 535-556.
- Seo, Y.-W., Giampapa, J. and Sycara, K. (2004) *Financial news analysis for intelligent portfolio management.*, Technical Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, Pittsburgh.
- Skinner, D. J. (1994) 'Why firms voluntarily disclose bad news', *Journal of Accounting Research*, 32(1), 38-60.
- Slattery, D. M., Sutcliffe, R. F. E. and Walsh, E. J. (2002) 'Automatic analysis of corporate financial disclosures', in Gillam, L. ed., *Making Money in the Financial Services Industry, a Workshop at the Terminology and Knowledge Engineering Conference (TKE 2002)*, Nancy, France, 30 August 2002.
- Sparck Jones, K. and Willett, P., eds. (1997). *Readings in information retrieval*, San Francisco, CA: Morgan Kaufmann.
- Stice, E. K. (1991) 'The market reaction to 10-K and 10-Q filings and to subsequent The Wall Street Journal earnings announcements', *The Accounting Review*, 66(1), 42-55.
- Subramaniam, R., Insley, R. G. and Blackwell, R. D. (1993) 'Performance and readability: A comparison of annual reports of profitable and unprofitable corporations', *The Journal of Business Communication*, 30(1), 49-61.

- Swales, G. S. and Yoon, Y. (1992) 'Applying artificial neural networks to investment analysis', *Financial Analysts Journal*, 48(5), 78-80.
- Tay, F. E. -H., Shen, L. and Cao, L. (2003) *Ordinary shares, exotic methods: Financial forecasting using data mining techniques*, Singapore: World Scientific.
- Tetlock, P. C. (2007) 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of Finance*, 62(3), 1139-1168.
- Tetlock, P. C. (2008) 'All the news that's fit to reprint: Do investors react to stale information?' *Working Paper*, Yale University.
- Tetlock, P. C., Saar-Tsechansky, M. and Macskassy, S. (2008) 'More than words: Quantifying language to measure firms' fundamentals', *The Journal of Finance*, 63(3), 1437-1467.
- Thomas, J. (1997) 'Discourse in the marketplace: The making of meaning in annual reports', *The Journal of Business Communication*, 34(1), 47-66.
- Thomas, J. D. (2003) *News and trading rules*, unpublished thesis (Ph.D.), Carnegie Mellon University.
- Thomas, J. D. and Sycara, K. (2000) 'Integrating genetic algorithms and text learning for financial prediction', in Freitas, A. A., Hart, W., Krasnogor, N. and Smith, J. eds., *Genetic and Evolutionary Computing 2000 Workshop on Data Mining with Evolutionary Algorithms*, Las Vegas, Nevada, 8-12 July 2000, 72-75.
- Thomsett, M. C. (2007) *The stock investor's pocket calculator: A quick guide to all the formulas and ratios you need to invest like a pro*, New York: AMACOM, American Management Association.
- Tumarkin, R. and Whitelaw, R. F. (2001) 'News or noise? Internet message board activity and stock prices', *Financial Analysts Journal*, 57(3), 41-51.
- van Bunningen, A. H. (2004) *Augmented trading: From news stories to stock price predictions using syntactic analysis*, unpublished thesis (M.Sc.), University of

Twente.

van Rijsbergen, C. J. (1979) *Information retrieval*, 2nd ed., London: Butterworths.

Whisenant, J. S., Sankaraguruswamy, S. and Raghunandan, K. (2003) 'Market reactions to disclosure of reportable events', *Auditing: A Journal of Practice and Theory*, 22(1), 181-194.

White, H. (1988) 'Economic prediction using neural networks: The case of IBM daily stock returns', *2nd Annual IEEE Conference on Neural Networks*, San Diego, 24-27 July 1988, IEEE, 451-458.

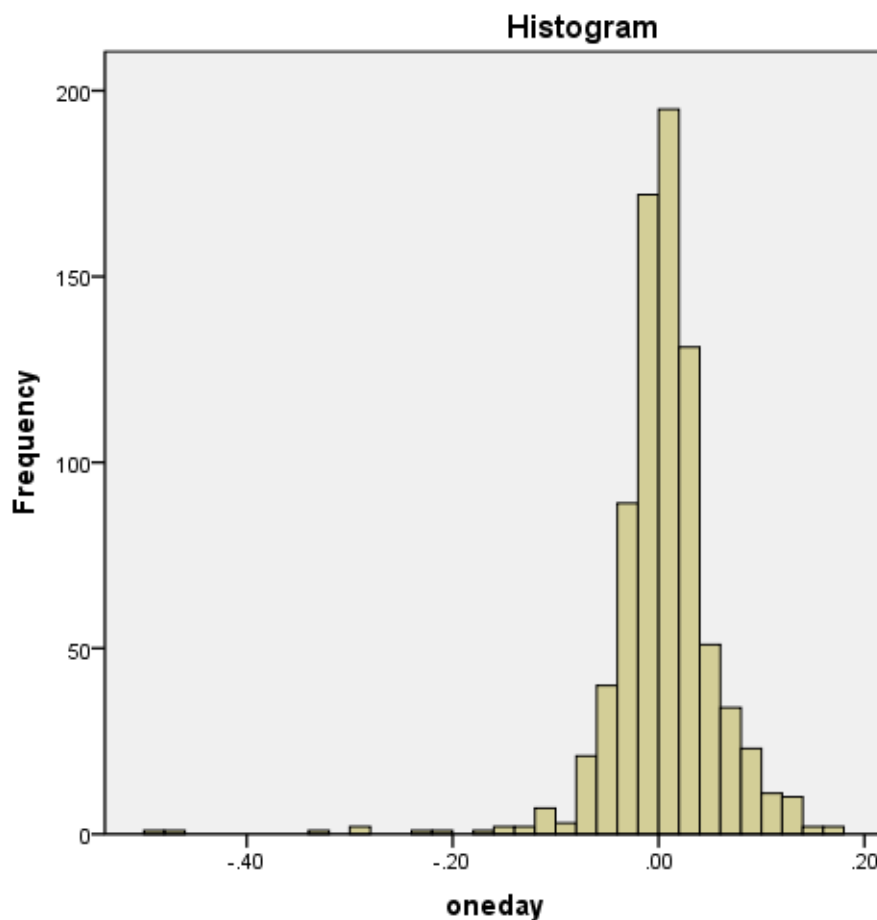
Wüthrich, B., Peramunetilleke, D., Leung, S., Cho, V., Zhang, J. and Lam, W. (1998) 'Daily prediction of major stock indices from textual WWW data', in Agrawal, R., Solorz, P. and Piatetsky, G. eds., *Fourth International Conference on Knowledge Discovery and Data Mining KDD-98*, New York, 27-31 August 1998, AAAI Press, 2720-2725.

Wysocki, P. D. (1999) *Cheap talk on the web: The determinants of postings on stock message boards*, Working Paper, University of Michigan Business School.

Yoo, P. D., Kim, M. H. and Jan, T. (2005) 'Financial forecasting: Advanced machine learning techniques in stock market analysis', *9th International Multitopic Conference (INMIC 2005)*, Karachi, Pakistan, 24-25 December 2005, IEEE.

## Appendix 1: Histogram for $t \pm 1$ days (1997–2000)

The histogram summarises  $\pm 1$  day results ( $n=803$ ). It is negatively skewed with a few unusual negative results. The mean and median are both close to zero (mean = -0.001 and median = 0). 90% of the data lie between  $0 \pm 0.1$ . The range is -0.49 to 0.17.



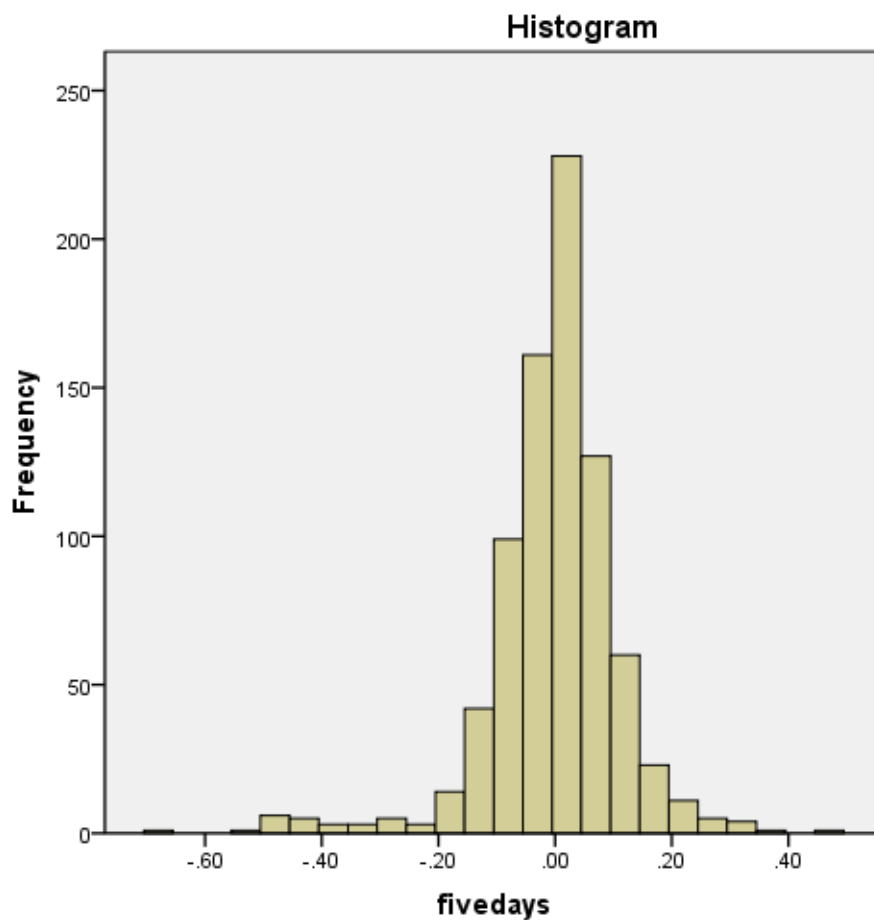
# Appendix 2: Boxplot for $t \pm 1$ days (1997–2000)

The box plot shows little variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



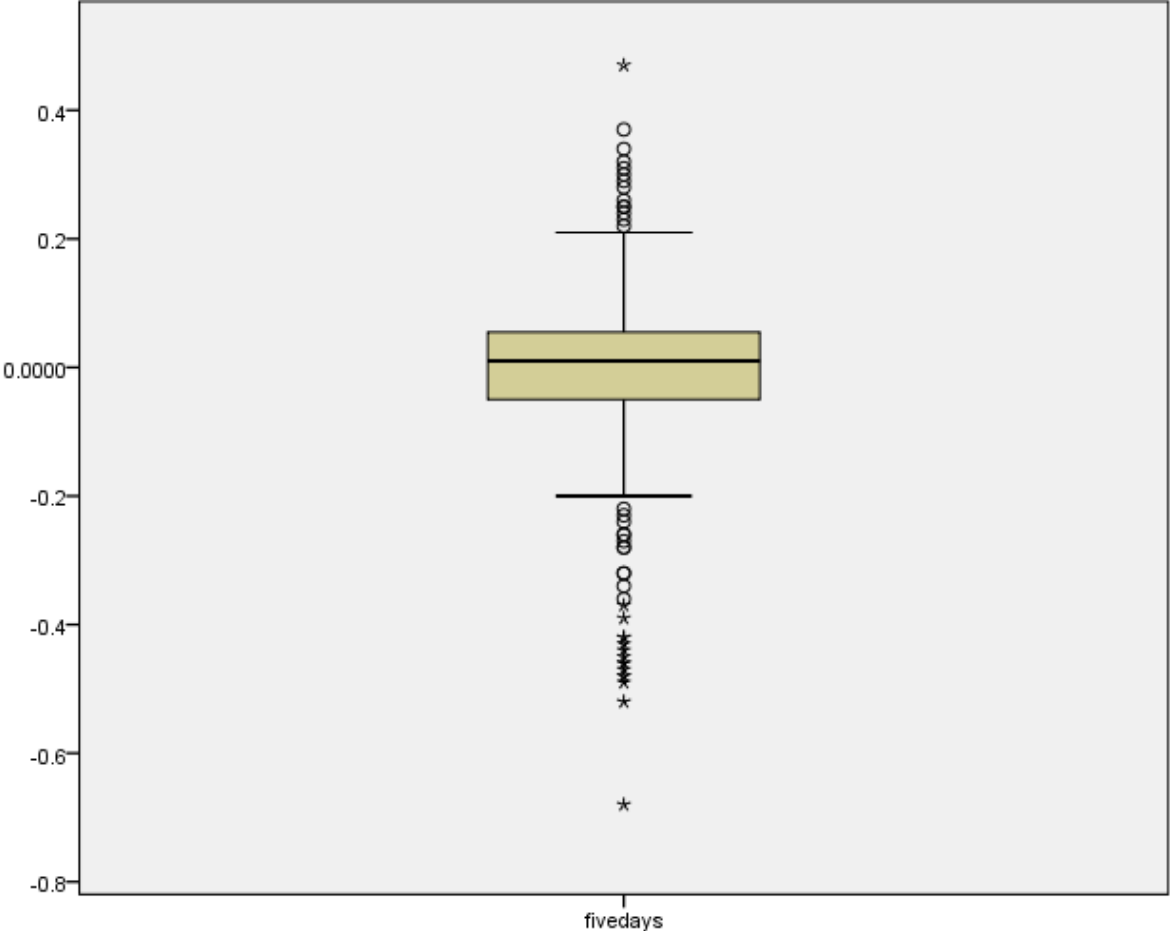
### Appendix 3: Histogram for $t \pm 5$ days (1997–2000)

The histogram below summarises  $\pm 5$  days results ( $n=803$ ). It is reasonably symmetric (slight negative skewness). The mean is  $-0.0012$  and the median =  $0.01$ . There is more variation in the results compared with  $\pm 1$  day. 90% of the values lie in the range  $0 \pm 0.15$ . The range is  $-0.68$  to  $0.47$ .



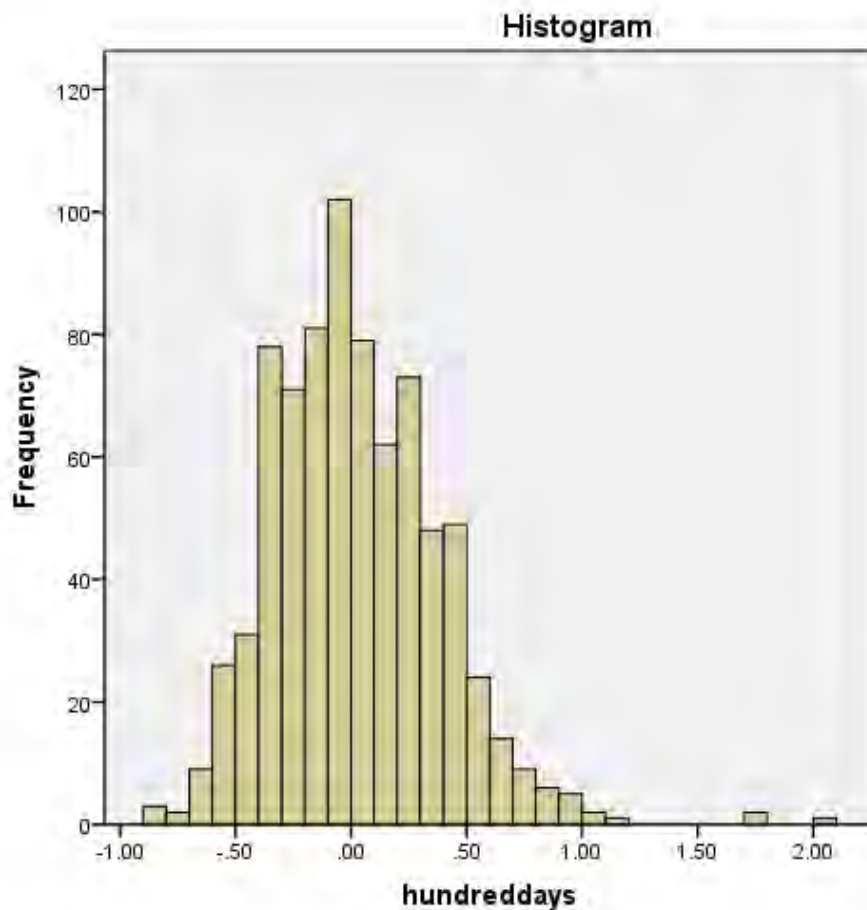
# Appendix 4: Boxplot for $t \pm 5$ days (1997–2000)

The box plot shows some variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



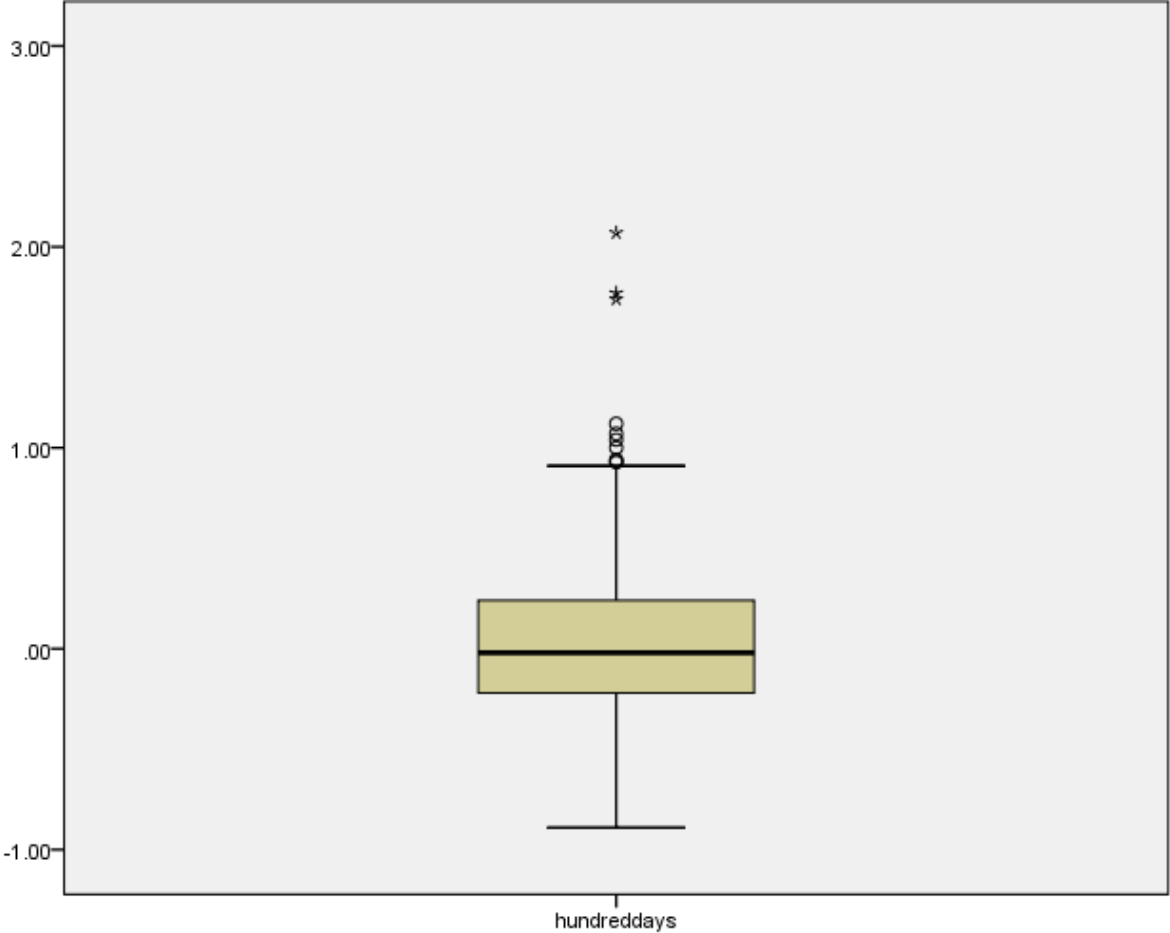
## Appendix 5: Histogram for $t \pm 100$ days (1997–2000)

The histogram below summarises  $\pm 100$  days results ( $n=778$ ). It is slightly positively skewed. The mean is 0.02 and the median = -0.02. There is considerably more variation than in the results for  $\pm 1$  day or  $\pm 5$  days. 90% of the values lie in the range  $0 \pm 0.60$ . The range is -0.89 to 2.07.



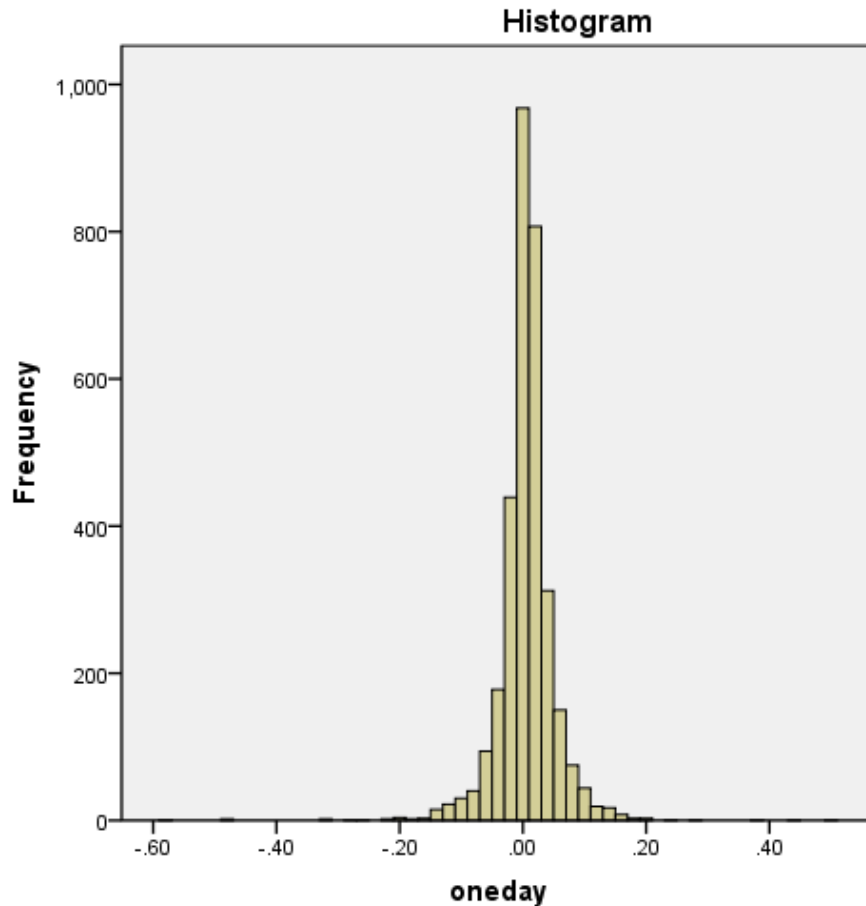
# Appendix 6: Boxplot for $t \pm 100$ days (1997–2000)

The box plot shows some variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



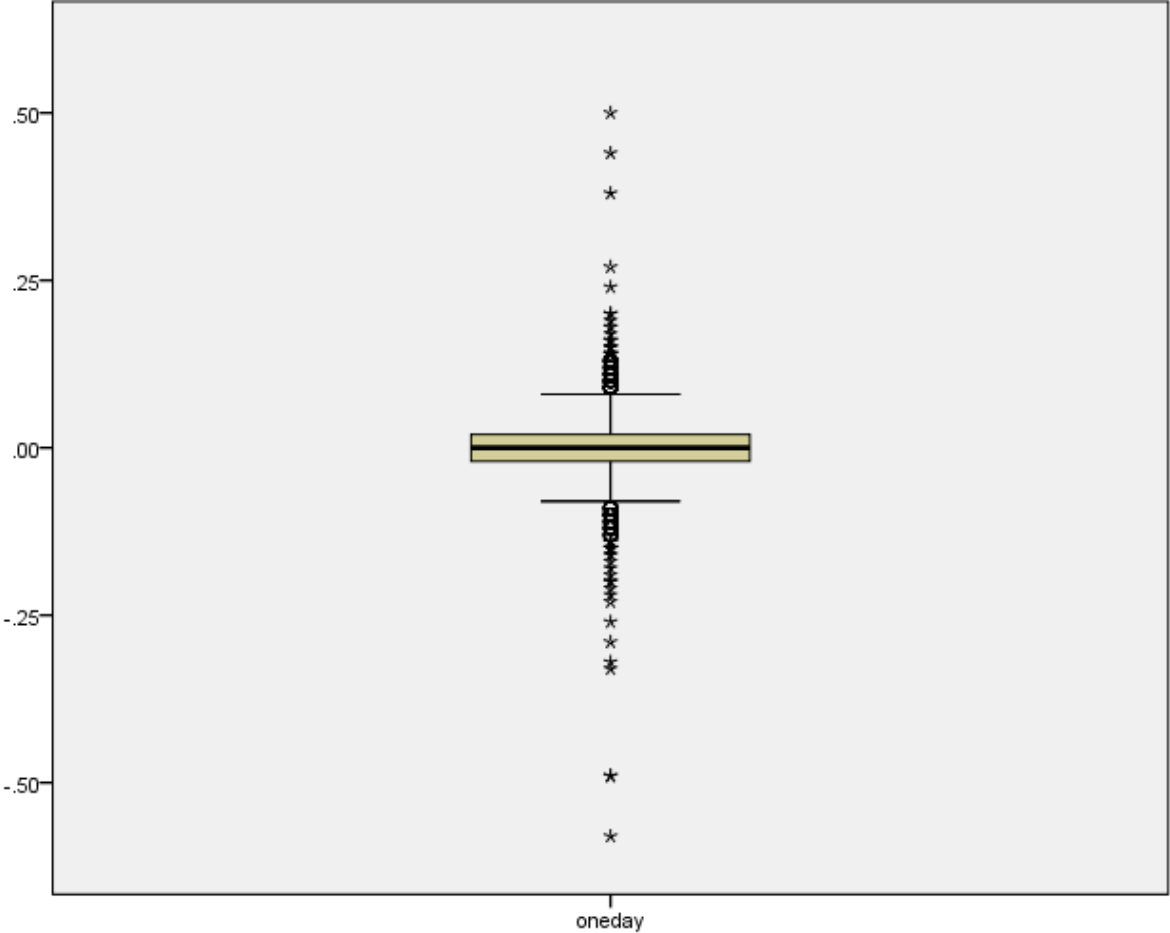
## Appendix 7: Histogram for $t \pm 1$ days (2005-2008)

The histogram summarises  $\pm 1$  day results ( $n=3247$ ). It is reasonably symmetric. The mean and median are both close to zero (mean = 0.001 and median = 0). 90% of the data lie between  $0 \pm 0.07$ . The range is -0.58 to 0.50.



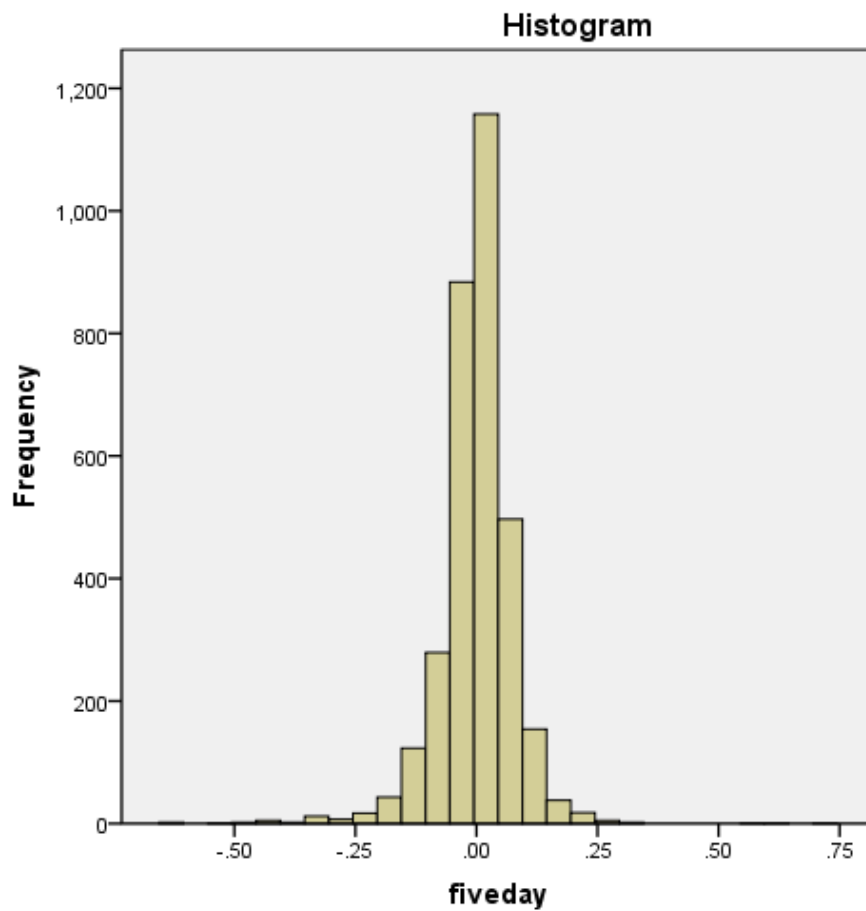
# Appendix 8: Histogram for $t \pm 1$ days (2005-2008)

The box plot shows little variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



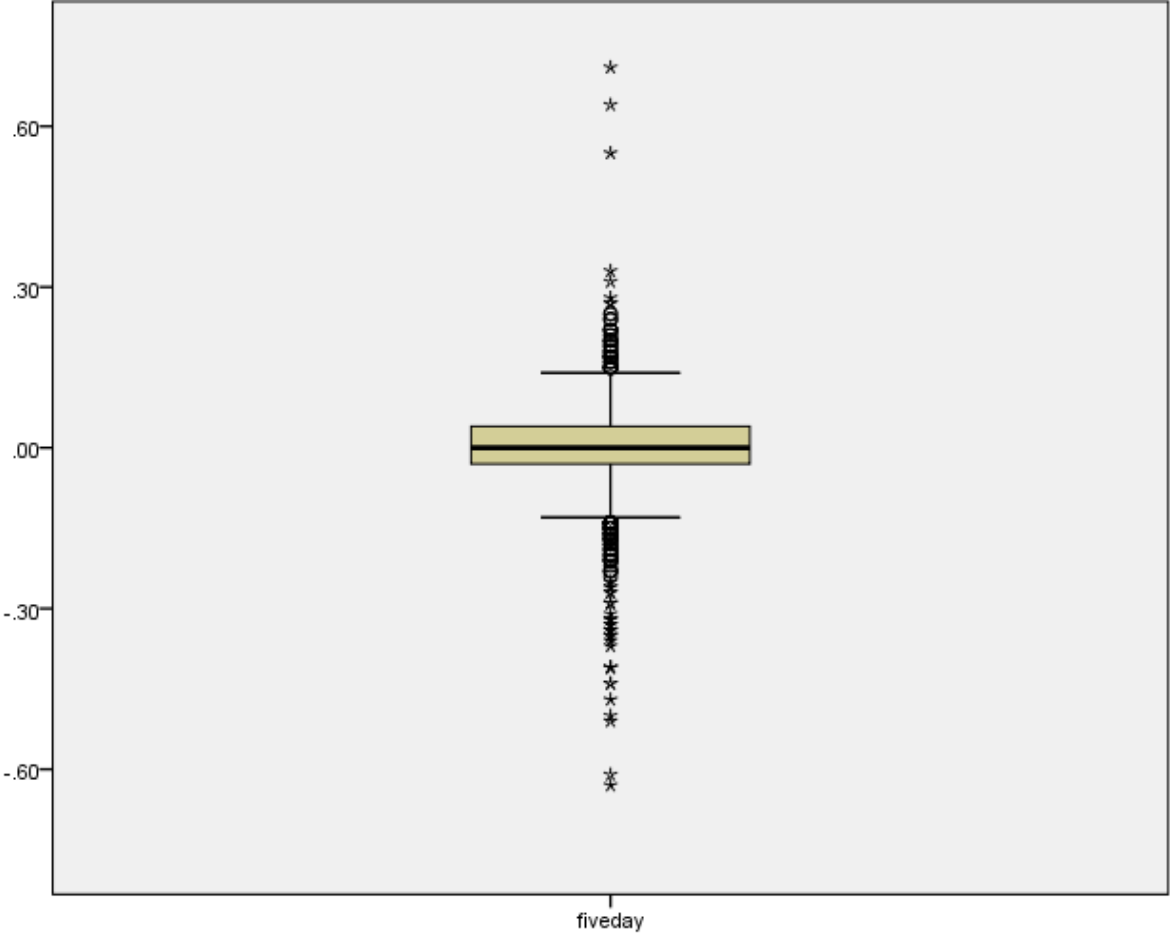
## Appendix 9: Histogram for $t \pm 5$ days (2005-2008)

The histogram below summarises  $\pm 5$  days results ( $n=3247$ ). It is reasonably symmetric. The mean is 0.0006 and the median is zero. There is more variation in the results compared with  $\pm 1$  day. 90% of the values lie in the range  $0 \pm 0.12$ . The range is -0.63 to 0.71.



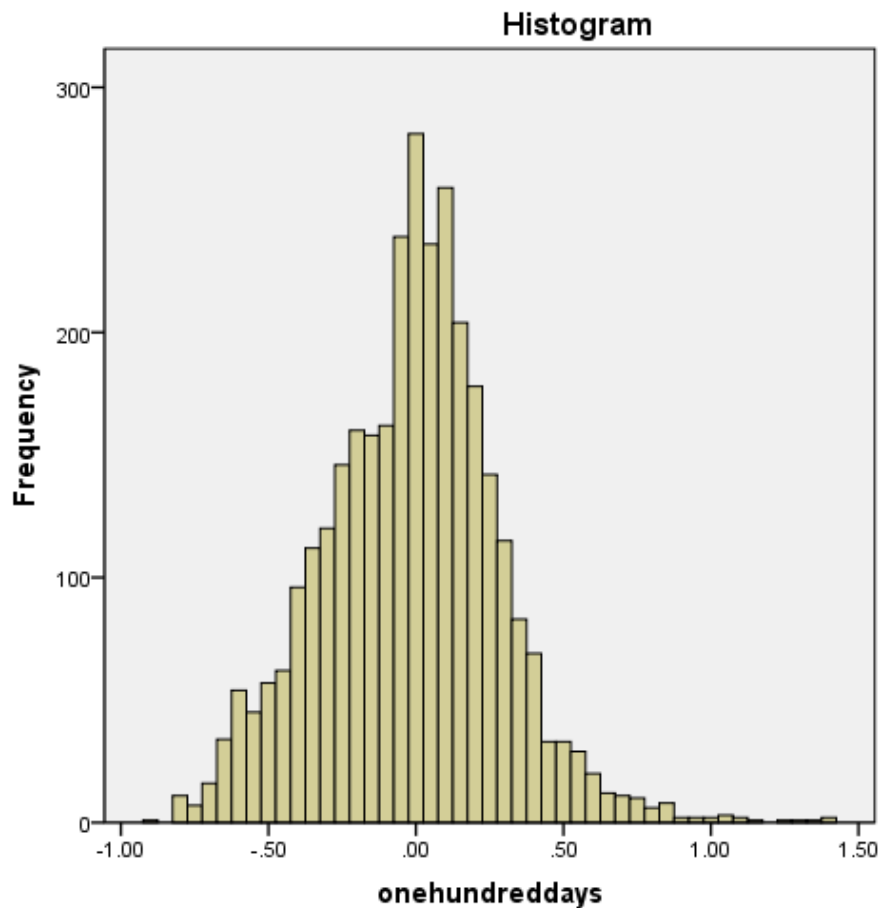
# Appendix 10: Boxplot for $t \pm 5$ days (2005-2008)

The box plot shows some variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



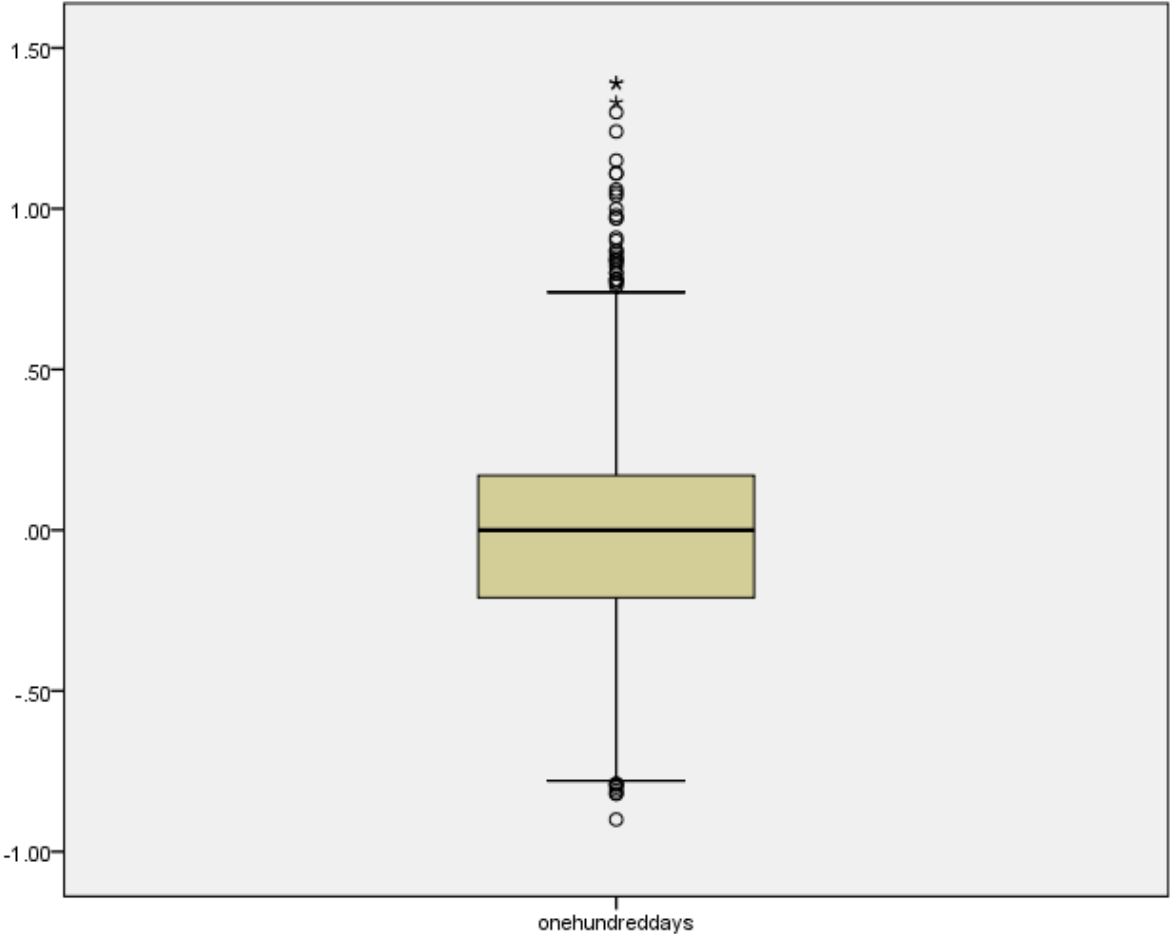
## Appendix 11: Histogram for $t \pm 100$ days (2005-2008)

The histogram below summarises  $\pm 100$  days results ( $n=3226$ ). It is reasonably symmetric. The mean is  $-0.016$  and the median is zero. There is considerably more variation than in the results for  $\pm 1$  day or  $\pm 5$  days. 90% of the values lie in the range  $0 \pm 0.50$ . The range is  $-0.9$  to  $1.39$ .

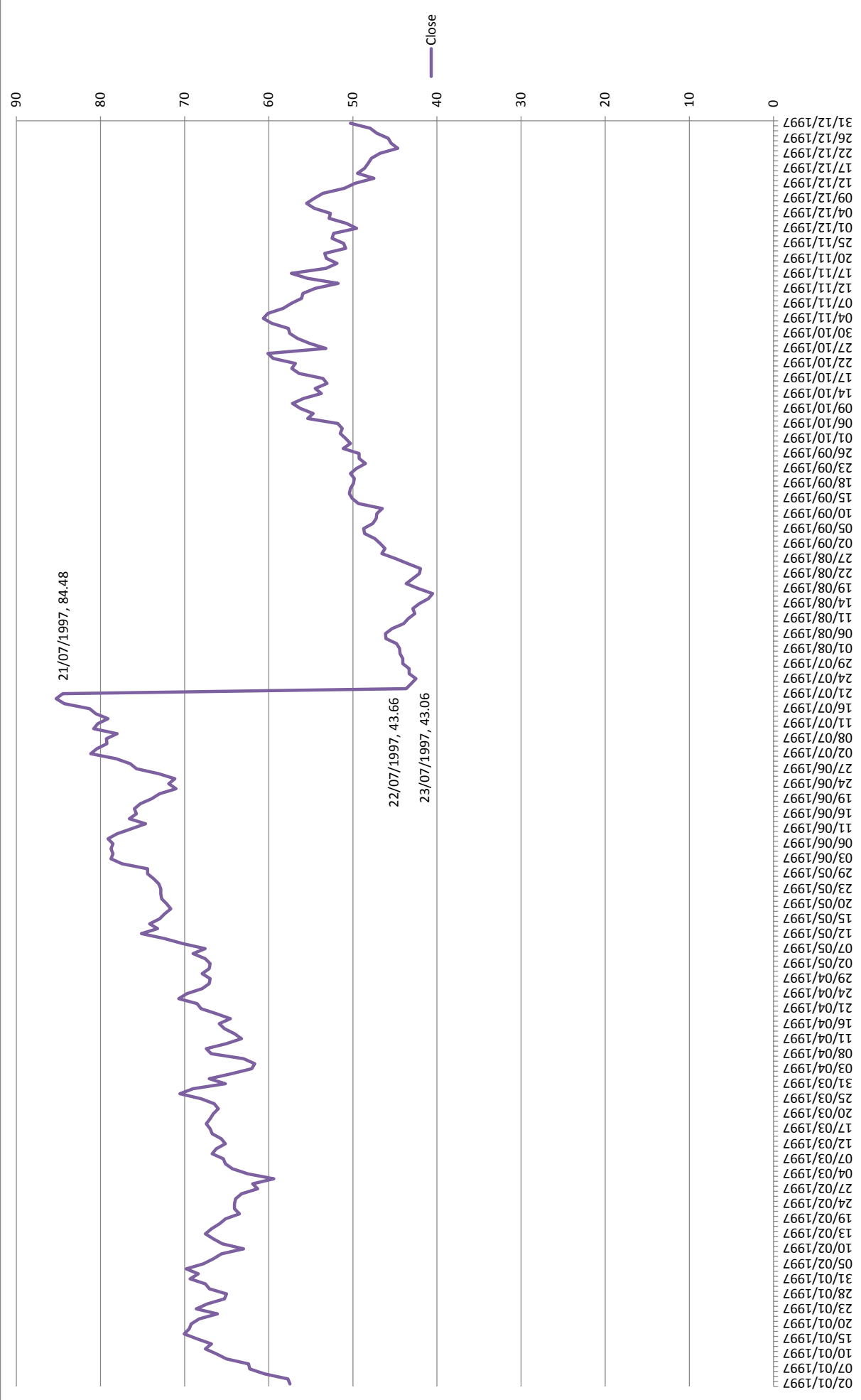


# Appendix 12: Boxplot for $t \pm 100$ days (2005-2008)

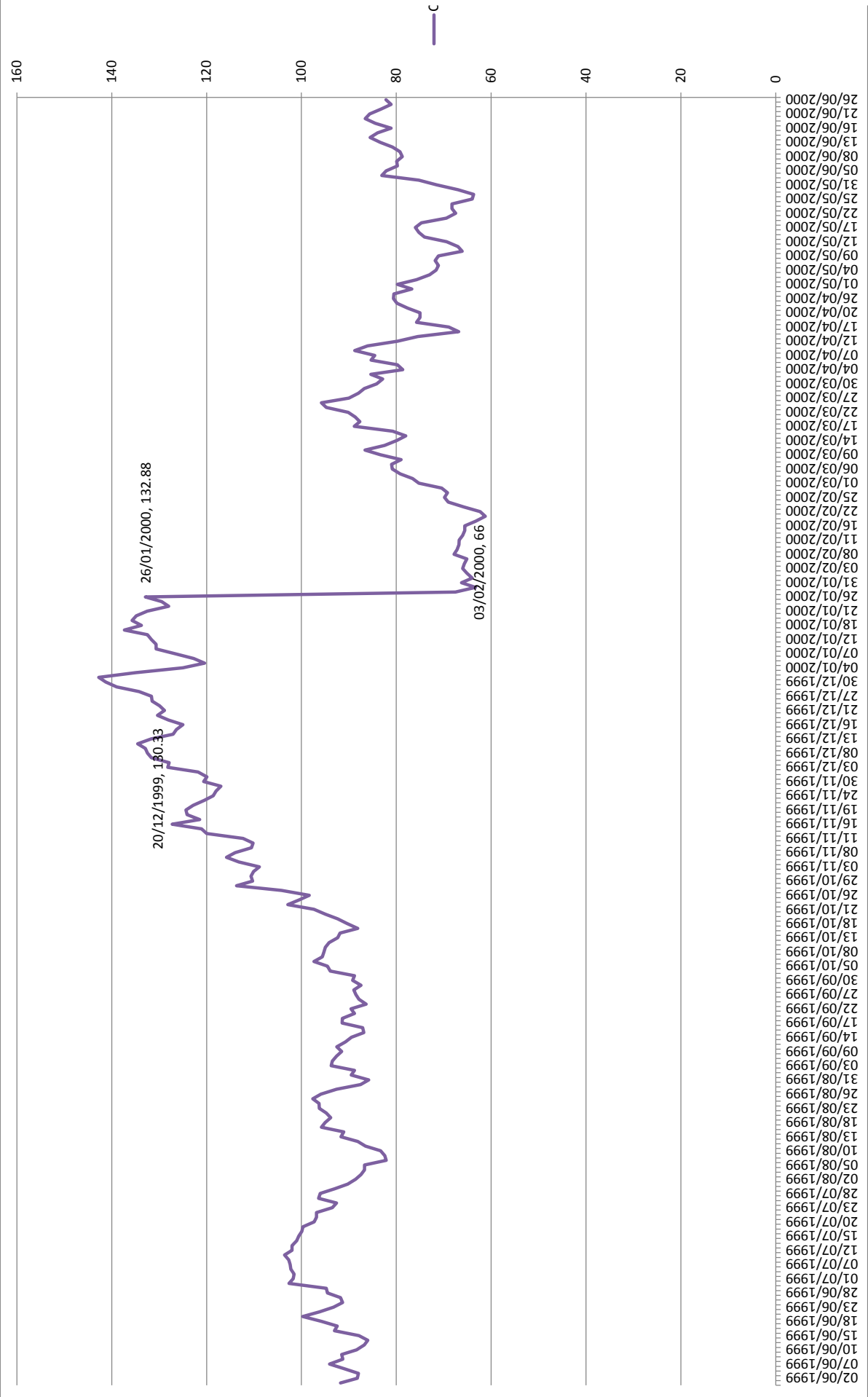
The box plot shows some variation around zero with some unusual values (circles represent outliers, asterisks extreme outliers).



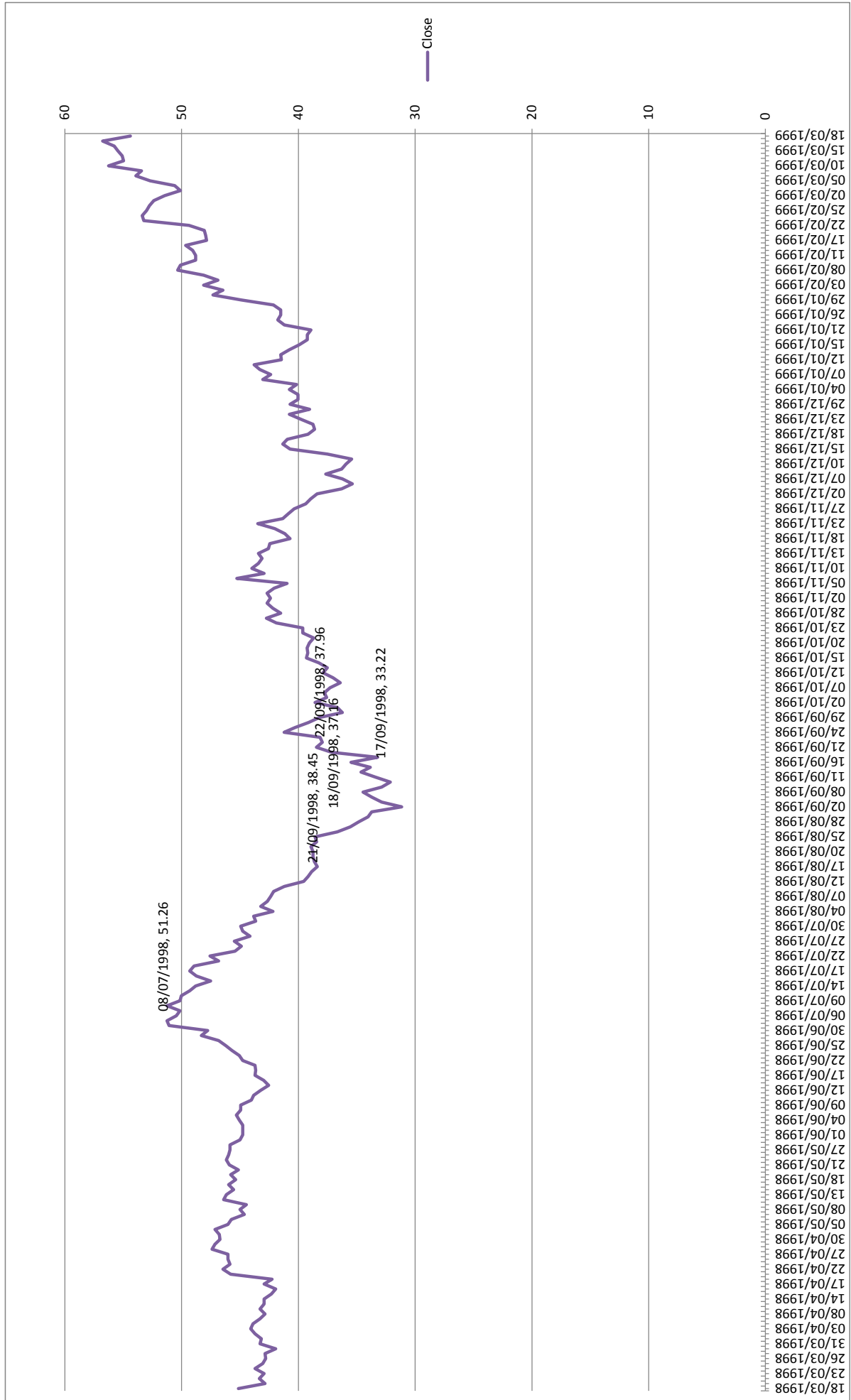
# Appendix 13: Time Series for Halliburton Co.



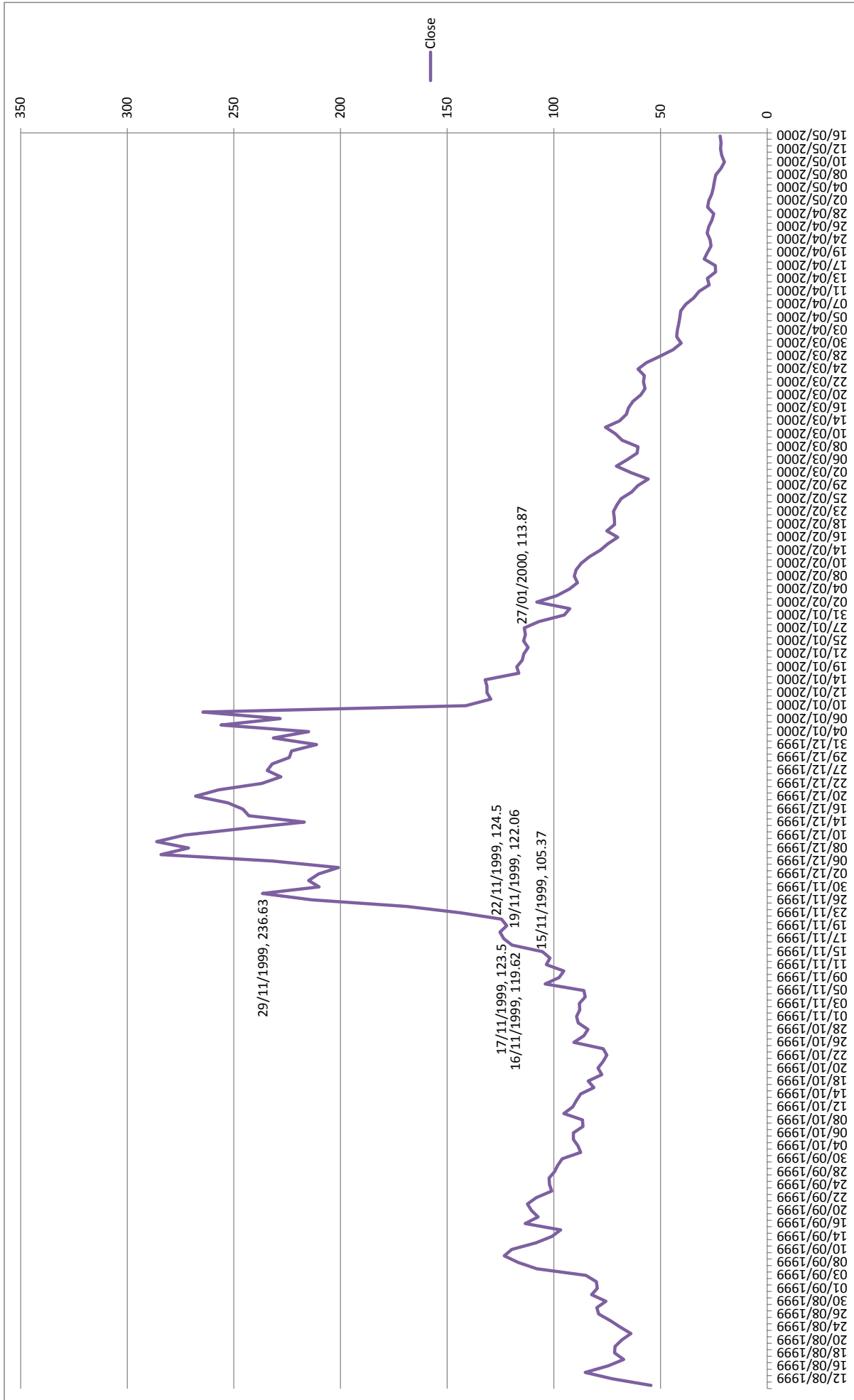
# Appendix 14: Time Series for Morgan Stanley Dean Witter & Co.



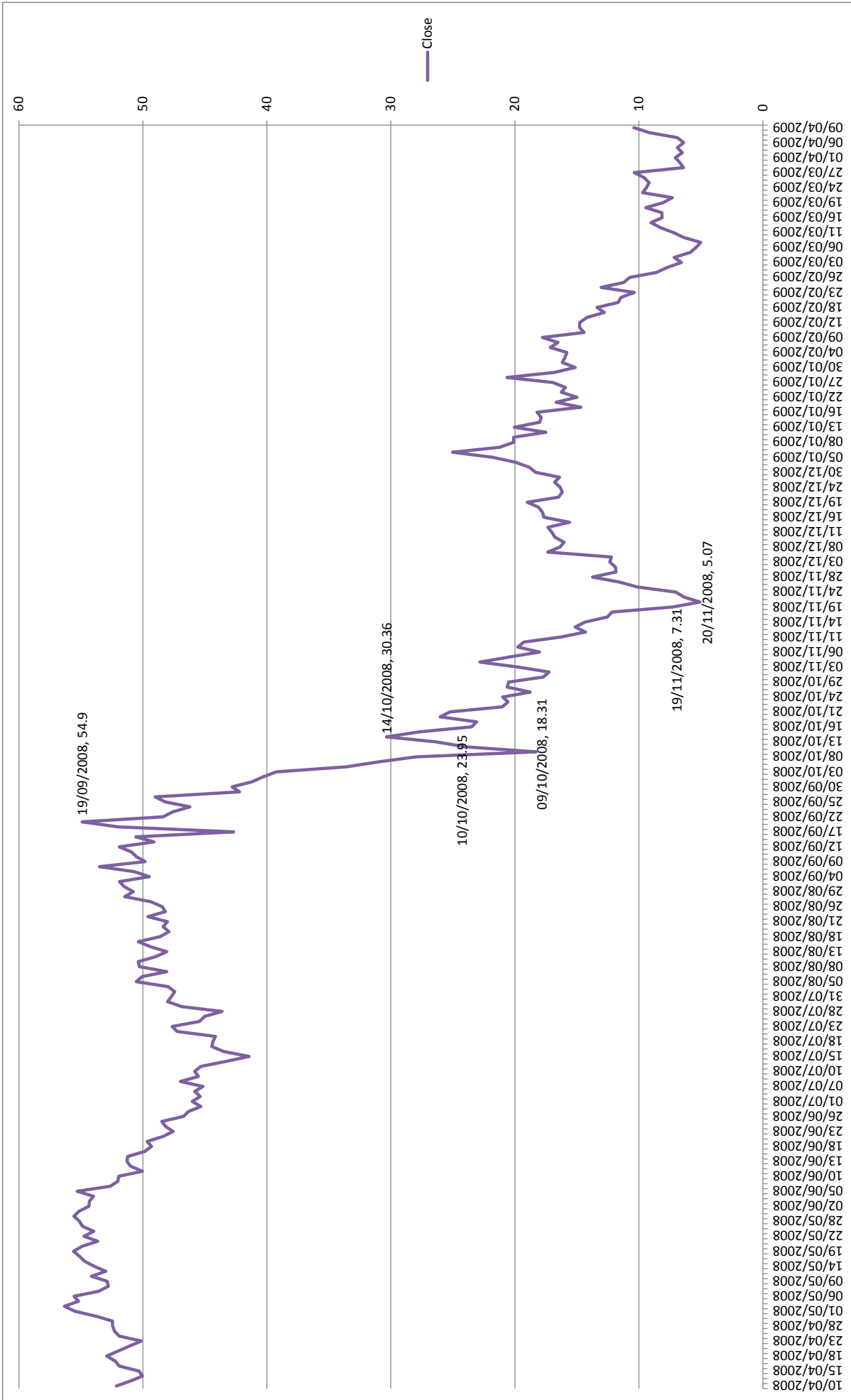
# Appendix 15: Time Series for Nike Inc.



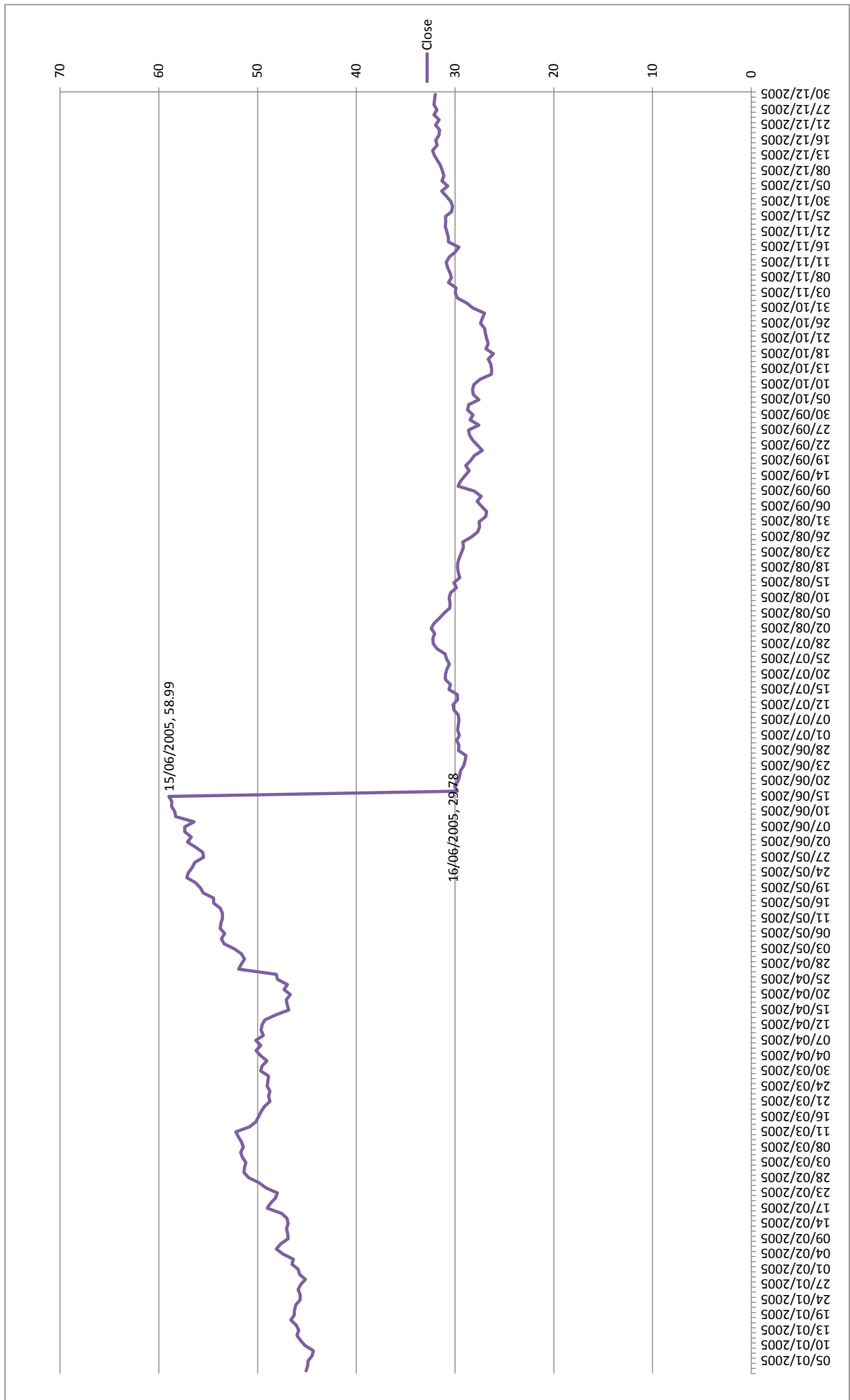
# Appendix 16: Time Series for Red Hat Inc.



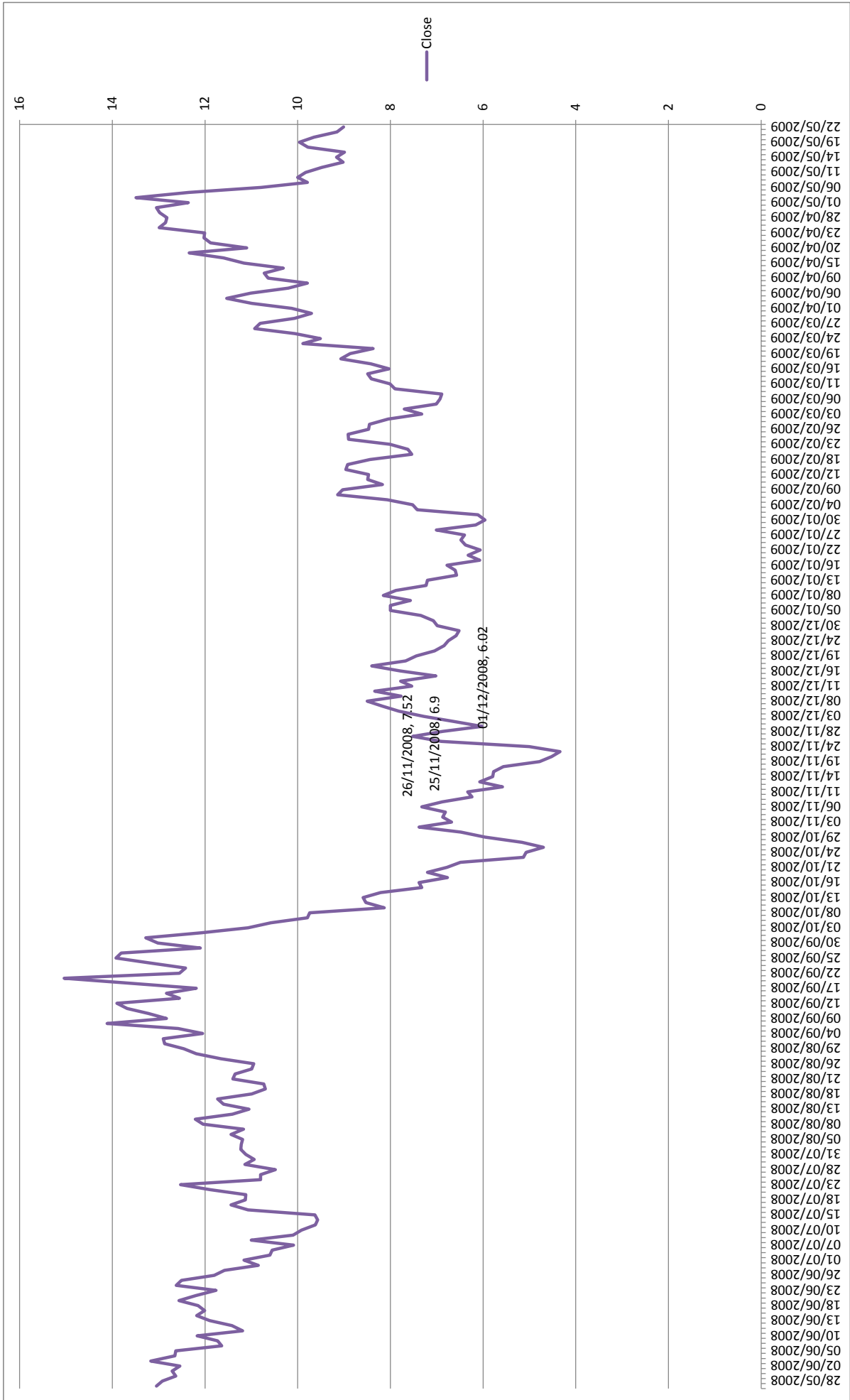
# Appendix 17: Time Series for Lincoln National Corp.



# Appendix 18: Time Series for O'Reilly Automotive Inc.



# Appendix 19: Time Series for Dr Horton Inc.



## **Appendix 20: Sample Phrases for Employment-Related FEPs**

### **fep\_accountant\_appointment**

- has dismissed the firm of (type\_accountant\_or\_accountant\_name)
- appointment as (something) principal accountants was terminated
- selected (something) to replace (type\_accountant\_or\_accountant\_name)
- has dismissed (type\_accountant\_or\_accountant\_name)

### **fep\_accountant\_appointment**

- engaged the accounting firm of (type\_accountant\_or\_accountant\_name)
- engaged (type\_accountant\_or\_accountant\_name)
- selected (something) (type\_accountant\_or\_accountant\_name) to replace

### **fep\_new\_personnel\_or\_promotions**

- will join (something) (type\_company\_or\_employee) as (type\_company\_or\_employee)
- appointed (something) as (type\_company\_or\_employee)
- plans to elect (something) as (type\_company\_or\_employee)
- will become (type\_company\_or\_employee)
- will take over from
- has been named (type\_company\_or\_employee)

### **fep\_potential\_new\_personnel**

- the board has retained an executive search firm to seek
- additions to the office of the president are also expected

### **fep\_remain\_as\_personnel**

- shall hold office until
- will continue to serve on
- will continue as (type\_company\_or\_employee)
- will remain actively involved

**fep\_resignation\_or\_leaving**

- company is willing to accept such resignation
- (type\_company\_or\_employee) will be leaving (something) in order to pursue other interests
- has resigned as the (type\_company\_or\_employee)
- resigned upon consummation of the merger
- resigned as the (type\_company\_or\_employee)

**fep\_layoff\_or\_dismiss\_personnel**

- any existing employment agreements with other employees of the company shall have been terminated
- eliminated approximately (something) positions related to
- is laying off approximately (something) employees

**fep\_potential\_employment\_problems**

- has at times experienced and continues to experience difficulty in recruiting qualified personnel
- loss of the services of any of the company's key employees could have a material adverse effect on
- there can be no assurance that the company will retain key (type\_company\_or\_employee) or attract such personnel in the future
- the loss of services of key personnel could have a material adverse effect on the company

# Appendix 21: List of FEP Types, Named Entities, and Types of Financial Object

## FEP Types

1. fep\_accountant\_dismissal
2. fep\_accountant\_appointment
3. fep\_new\_personnel\_or\_promotions
4. fep\_potential\_new\_personnel
5. fep\_remain\_as\_personnel
6. fep\_resignation\_or\_leaving
7. fep\_layoff\_or\_dismiss\_personnel
8. fep\_potential\_employment\_problems
9. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
10. fep\_amend\_agreement\_and\_plan\_of\_merger
11. fep\_close\_agreement\_and\_plan\_of\_reorganization
12. fep\_securities\_purchase\_agreement
13. fep\_stock\_option\_agreement\_or\_plan
14. fep\_rights\_plan\_issue\_offer\_or\_agreement
15. fep\_amend\_rights\_plan\_issue\_offer\_or\_agreement
16. fep\_stock\_offering
17. fep\_private\_placement
18. fep\_cancel\_private\_placement\_of\_stock
19. fep\_dividend\_distribution
20. fep\_indenture
21. fep\_purchase\_agreement
22. fep\_amend\_purchase\_agreement
23. fep\_product\_offering
24. fep\_cancel\_or\_postpone\_stock\_offering
25. fep\_stock\_or\_note\_conversion
26. fep\_stock\_split
27. fep\_reverse\_stock\_split
28. fep\_will\_or\_has\_generated\_revenue
29. fep\_will\_or\_has\_generated\_revenue\_decrease

30. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill
31. fep\_no\_goodwill
32. fep\_will\_or\_has\_generated\_loss
33. fep\_will\_or\_has\_generated\_income
34. fep\_will\_or\_has\_generated\_income\_decrease
35. fep\_adopt\_strategy\_or\_plan
36. fep\_has\_or\_will\_adopt\_financial\_standard
37. fep\_adopt\_resolution\_to\_designate\_or\_create\_stock
38. fep\_copyright\_infringement
39. fep\_counterclaim
40. fep\_meets\_or\_expects\_to\_meet\_listing\_requirements
41. fep\_hasnt\_or\_mightnt\_meet\_listing\_requirements
42. fep\_granted\_an\_exception
43. fep\_change\_state\_of\_incorporation
44. fep\_relocation
45. fep\_could\_would\_can\_expect\_material\_adverse\_effect
46. fep\_could\_would\_can\_not\_expect\_material\_adverse\_effect
47. fep\_breach
48. fep\_owes\_compensation
49. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info

**Named entities and types of financial object**

1. type\_accountant\_or\_accountant\_name
2. type\_appointment\_or\_promotion
3. type\_company\_or\_employee
4. type\_resignation\_or\_departure
5. type\_stock\_option\_agreement\_or\_plan
6. type\_rights\_plan\_issue\_offer\_or\_agreement
7. type\_private\_placement\_of\_stock
8. type\_stock\_purchase\_agreement
9. type\_public\_stock\_offering
10. type\_stock\_right\_or\_option
11. type\_purchase\_agreement

12. type\_acquisition\_or\_merger
13. type\_loss
14. type\_income
15. type\_revenue
16. type\_stock\_market
17. type\_gen\_ag\_report\_doc\_or\_info
18. type\_valid\_year

## Appendix 22: List of Keywords

ability	act	affairs
absence	acting	affect
absolute	action	affected
accelerate	actions	affecting
acceleration	activities	affiliate
accept	activity	affiliated
acceptable	acts	affiliates
acceptance	actual	agency
accepted	added	agent
access	addition	agents
accession	additional	aggregate
accompanying	address	agree
accordance	addressed	agreed
account	addresses	agreement
accountants	adequate	agreements
accounted	adjusted	agrees
accounting	adjustment	alleged
accounts	adjustments	allocated
accrued	administration	allocation
accumulated	administrative	allowance
accuracy	adopt	alter
accurate	adopted	alternative
acknowledge	adoption	amend
acknowledged	advance	amended
acknowledges	advances	amendment
acquire	adverse	amendments
acquired	adversely	amortization
acquiring	advice	amount
acquiror	advisable	amounts
acquisition	advised	analysis
acquisitions	advisors	annex

anniversary	assessments	automatically
announced	asset	availability
announcement	assets	avenue
annual	assign	average
anticipated	assigned	award
applicable	assigning	aware
application	assignment	balance
applications	assigns	balances
applied	assistant	bank
apply	associate	banking
appoint	associates	bankruptcy
appointed	association	bargaining
appointment	assume	base
appraisal	assumed	based
approval	assuming	basic
approvals	assumption	basis
approve	assurance	bear
approved	assurances	begin
approximately	attached	beginning
arbitration	attention	behalf
arbitrator	attorney	believes
arbitrators	attorneys	beneficial
area	attributable	beneficially
arise	audit	benefit
arises	audited	benefits
arising	auditors	best
arrangement	audits	bet
arrangements	authorities	bid
article	authority	binding
articles	authorization	board
ascribed	authorizations	boards
asserted	authorize	body
assessment	authorized	bond

bonds	caused	collective
bonus	cease	collectively
book	central	combination
books	ceo	combined
borne	certificate	commenced
borrower	certificates	commencement
bound	certification	commencing
breach	certified	commerce
breached	chain	commercial
breaches	chairman	commercially
broker	change	commission
brokerage	changed	commissions
brokers	character	commitment
brought	charge	commitments
business	charged	committee
businesses	charges	common
buyer	charter	communication
bylaws	chief	communications
calculated	circumstance	companies
calendar	circumstances	company
call	claim	comparable
called	claims	compared
canceled	class	compensation
cancellation	classification	compete
capacity	clause	competent
capital	clauses	competition
capitalization	clear	competitive
capitalized	client	complete
carry	close	completed
carrying	closing	completion
case	code	compliance
cash	collateral	complied
cause	collection	comply

compromise	consultants	copyrights
computation	consulting	corporate
computer	consummate	corporation
concurrently	consummated	corporations
condensed	consummation	correct
condition	contact	cost
conditions	contained	costs
conduct	contemplated	counsel
conducted	contents	count
confidential	contingent	counterparts
confidentiality	continue	courier
conflict	continued	course
conflicts	continuing	court
conformed	contract	covenant
conformity	contracts	covenants
connection	contractual	coverage
consecutive	contrary	covered
consent	contributed	covering
consents	contribution	create
consequences	contributions	created
consideration	control	creation
considered	controlled	credit
consist	controlling	creditors
consistent	controls	cumulative
consists	conversion	currency
consolidated	convert	current
consolidation	converted	custodian
constitute	convertible	customary
constitutes	conveyance	customer
constituting	cooperate	customers
construction	copies	damage
construed	copy	damages
consultant	copyright	data

database	demands	discharge
date	department	disclose
dated	deposit	disclosed
dates	depository	disclosure
debenture	deposited	disclosures
debt	depreciation	discretion
debtor	described	discussed
debts	description	discussions
decision	design	disposal
declaration	designate	dispose
declare	designated	disposed
declared	designation	disposition
decrease	designations	dispute
decree	designed	disputes
deed	desirable	dissenting
deem	destroyed	dissolution
deemed	destruction	distribute
default	determination	distributed
defaults	determine	distribution
defend	determined	distributions
defending	determines	distributor
defense	determining	distributors
deferred	develop	district
deficit	developed	dividend
defined	development	dividends
definitions	differ	dividing
definitive	diluted	division
delay	direct	document
deliver	directly	documentation
delivered	director	documents
delivering	directors	dollar
delivery	disability	dollars
demand	disbursements	domestic

drive	enforceable	events
due	enforced	evidence
duly	enforcement	evidenced
duties	engage	evidencing
duty	engaged	exact
earlier	engineering	exceed
earliest	enhanced	exceeds
earnings	enter	exceptions
economic	entered	excess
effect	enterprise	exchange
effected	entire	exchangeable
effective	entirety	exchanged
effectiveness	entities	excluded
effects	entitled	excluding
efforts	entity	exclusive
elect	entry	execute
elected	environment	executed
election	environmental	execution
electronic	equal	executive
eligible	equipment	exempt
employed	equitable	exemption
employee	equity	exercisable
employees	equivalent	exercise
employer	equivalents	exercised
employment	erudite	exhibit
enable	escrow	exhibits
encumbrance	escrowed	exist
encumbrances	establish	existence
end	established	existing
ended	estimated	exists
ending	estimates	expected
enforce	europe	expects
enforceability	event	expenditures

expense	filer	gaap
expenses	filing	gain
experience	filings	general
expiration	film	generality
expire	final	generally
expired	finance	genius
express	financial	give
expressed	financing	giving
expressly	firm	good
extend	fiscal	goods
extended	fixed	goodwill
extension	floor	governed
extent	flows	governing
facilitate	force	government
facilities	foregoing	governmental
facility	foreign	grant
fact	form	granted
factors	forma	grantee
facts	forms	grantor
fails	forward	grants
failure	fraction	greater
fair	fractional	gross
fairly	franchise	group
faith	franchises	growth
family	free	guarantee
favor	frontline	guarantor
favorable	full	guaranty
federal	fully	hand
fee	fund	hardware
fees	funds	harmless
fiduciary	furnish	hazardous
file	furnished	headings
filed	future	health

held	incurred	intended
high	indebtedness	intends
historical	indemnification	intent
hold	indemnified	intention
holder	indemnify	interactive
holders	indemnifying	interest
holding	indemnitee	interests
holdings	indemnity	interfere
holds	indenture	interim
hours	independent	internal
hundred	index	international
identification	indirect	internet
identified	indirectly	interpretation
illegal	individual	inure
immediately	individually	invalid
impact	industrial	inventory
impair	industry	investigation
imposed	information	investment
improvements	infringement	investments
inaccuracy	initial	investor
incentive	injunction	investors
include	inquiry	involve
included	insiders	involved
includes	insolvency	involving
including	instructions	irrevocable
inclusion	instrument	irrevocably
income	instruments	irs
incorporated	insurance	issuable
incorporation	intangible	issuance
increase	integral	issue
increased	integrated	issued
increases	integration	issuer
incur	intellectual	issues

item	license	maintain
items	licensed	maintained
joint	licenses	maintenance
jointly	licensing	major
judgment	lien	majority
july	liens	make
junior	lieu	makes
jurisdiction	life	making
jurisdictions	light	management
key	limit	manager
kind	limitation	managing
knowledge	limitations	mandatory
labor	limited	manner
language	limiting	manufacturing
lapse	line	mark
law	liquidation	market
laws	list	marketable
leading	listed	marketing
lease	listing	markets
leased	lists	marks
leasehold	litigation	material
leases	loan	materially
legal	loans	materials
legally	local	matter
legend	located	matters
lender	long	maximum
lesser	longer	meaning
letter	loss	meanings
letters	losses	means
level	lost	meeting
liabilities	made	meetings
liability	mail	member
liable	mailed	members

memorandum	net	offices
merge	network	official
merged	nonassessable	omision
merger	normal	omission
message	note	omissions
method	notes	omit
methods	notice	omitted
milestone	notices	open
million	notification	operate
minimum	notified	operated
miscellaneous	notify	operating
misleading	notwithstanding	operation
misrepresentation	number	operations
modification	numbers	opinion
modified	object	opinions
modify	obligated	opportunity
money	obligation	option
month	obligations	options
months	obtain	oral
mortgage	obtained	order
multiplied	obtaining	orders
multiplying	occur	ordinary
mutual	occurred	organization
mutually	occurrence	organized
named	occurring	original
names	occurs	originator
nasdaq	offer	outstanding
national	offered	overnight
nature	offering	owned
nearest	offers	owner
necessary	office	ownership
negotiation	officer	owns
negotiations	officers	page

paid	period	practices
par	periods	pre
paragraph	permit	precedent
parent	permits	preceding
park	permitted	preemptive
part	person	preference
participate	personal	preferences
participating	personally	preferred
participation	personnel	preliminary
parties	persons	premises
partner	pertaining	premium
partners	phone	prepackaged
partnership	place	prepaid
party	placement	preparation
past	plan	prepare
patent	planning	prepared
patents	plans	present
pay	platforms	presented
payable	platinum	presently
paying	pledge	preserve
payment	policies	president
payments	policy	press
payroll	pooling	prevent
pc	portion	previously
penalties	position	price
penalty	possession	prices
pending	post	primarily
pension	postage	primary
percent	potential	principal
percentage	power	principles
perform	powers	prior
performance	practicable	privacy
performed	practice	private

privileges	provider	reason
pro	providing	reasonable
procedures	provision	reasonably
proceeding	provisions	recapitalization
proceedings	proxies	receipt
proceeds	proxy	receipts
process	public	receivable
produce	publicly	receivables
product	published	receive
products	purchase	received
professional	purchased	receives
profit	purchaser	receiving
profits	purchasers	recent
program	purchases	recitals
programs	purchasing	reclassification
prohibited	purpose	recognition
prompt	purposes	recognized
promptly	pursuant	recommendation
promulgated	qualification	record
proper	qualifications	recorded
properly	qualified	records
properties	qualify	red
property	quarter	redeem
proportion	quarterly	redeemable
proposal	quorum	redeemed
proposed	quotation	redemption
proprietary	rata	reduce
prospects	rate	reduced
prospectus	rates	reduction
protect	ratio	refer
protection	rational	reference
provide	read	references
provided	real	referred

reflect	remainder	resolution
reflected	remaining	resolutions
refusal	remedies	resolved
regard	remedy	resource
register	removal	resources
registered	reorganization	respect
registrable	report	respective
registrant	reported	respects
registration	reporting	response
registrations	reports	responsibility
regular	represent	responsible
regulation	representation	restated
regulations	representative	restricted
regulatory	represented	restriction
reimburse	representing	restrictions
relate	represents	restrictive
related	repurchase	restructuring
relates	request	result
relating	requested	resulted
relation	requests	resulting
relations	require	results
relationship	required	retain
relationships	requirement	retained
relative	requirements	retirement
release	requires	return
released	requiring	returns
releases	requisite	revenue
relevant	resale	revenues
reliance	research	reverse
relief	reserve	review
relieve	reserved	right
rely	reserves	rights
remain	resignation	rise

risk	seller	single
risks	sellers	site
road	selling	software
rounded	senior	sold
royalties	sentence	sole
royalty	separate	solely
rule	series	solicit
rules	serve	solicitation
safety	server	solution
salary	service	solutions
sale	services	soon
sales	set	sought
satisfaction	sets	source
satisfactory	setting	south
satisfied	settle	special
satisfy	settlement	specific
schedule	severability	specifically
schedules	severally	specified
scope	severance	split
sec	share	standard
secret	shareholder	standards
secretary	shareholders	standing
secrets	shares	state
section	sharing	stated
sections	sheet	statement
secure	sheets	statements
securities	short	states
security	shown	stating
securityholder	signature	status
seek	signatures	statute
seeking	signed	statutes
selected	significant	statutory
sell	similar	sterling

stock	superior	telecopy
stockholder	supplement	telephone
stockholders	supplemental	template
stolen	supplemented	temporary
stop	supplements	tender
storage	supplied	term
strategic	supplier	terminate
strategy	suppliers	terminated
sub	supply	termination
subdivision	support	terms
subject	surrender	test
sublicense	surrendered	third
submission	survival	threatened
submitted	survive	time
subscription	surviving	timely
subsection	suspension	times
subsequent	system	title
subsidiaries	systems	titles
subsidiary	table	tm
subsidies	takeover	today
substance	taking	tools
substantial	tangible	total
substantially	target	trade
successor	tax	traded
successors	taxable	trademark
suffered	taxation	trademarks
sufficient	taxes	trading
suit	taxing	transaction
suite	tech	transactions
suits	technical	transfer
sum	technological	transferee
summary	technologies	transferred
summit	technology	transfers

transitory	users	warrant
transom	valid	warranties
treasurer	validity	warrants
treasury	validly	warranty
treated	valuation	web
treatment	value	webmaster
trust	values	weighted
trustee	vendor	whole
type	vendors	wholly
unable	venture	wide
unaudited	ventures	willful
uncertainties	versions	winding
undersigned	vest	windows
understanding	vested	withdrawn
understandings	vesting	withheld
understands	vice	withholding
understood	view	witness
undertaking	violate	words
underwriter	violation	work
underwriters	violations	working
underwriting	virtue	works
underwritten	visual	world
unenforceable	void	worldwide
union	vote	writ
unit	votes	write
united	voting	writing
unpaid	waive	written
unreasonably	waived	year
untrue	waiver	years
user	waivers	

## **Appendix 23: Different FEP Types Recognised in the First S&P 500 Dataset**

Downs:

1. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
2. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill
3. fep\_could\_would\_can\_expect\_material\_adverse\_effect
4. fep\_indenture
5. fep\_new\_personnel\_or\_promotions
6. fep\_purchase\_agreement
7. fep\_resignation\_or\_leaving
8. fep\_rights\_plan\_issue\_offer\_or\_agreement
9. fep\_securities\_purchase\_agreement
10. fep\_stock\_offering
11. fep\_stock\_option\_agreement\_or\_plan
12. fep\_stock\_or\_note\_conversion
13. fep\_stock\_purchase\_agreement
14. fep\_stock\_split
15. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
16. fep\_will\_or\_has\_generated\_income
17. fep\_will\_or\_has\_generated\_loss
18. fep\_will\_or\_has\_generated\_revenue

Ups:

1. fep\_accountant\_appointment
2. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
3. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill
4. fep\_amend\_rights\_plan\_issue\_offer\_or\_agreement
5. fep\_could\_would\_can\_expect\_material\_adverse\_effect
6. fep\_indenture
7. fep\_new\_personnel\_or\_promotions
8. fep\_purchase\_agreement
9. fep\_remain\_as\_personnel
10. fep\_resignation\_or\_leaving
11. fep\_rights\_plan\_issue\_offer\_or\_agreement
12. fep\_securities\_purchase\_agreement
13. fep\_stock\_or\_note\_conversion
14. fep\_stock\_purchase\_agreement
15. fep\_stock\_split
16. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
17. fep\_will\_or\_has\_generated\_income
18. fep\_will\_or\_has\_generated\_loss
19. fep\_will\_or\_has\_generated\_revenue

## Downs and Ups:

1. fep\_accountant\_appointment
2. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
3. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill
4. fep\_amend\_rights\_plan\_issue\_offer\_or\_agreement
5. fep\_could\_would\_can\_expect\_material\_adverse\_effect
6. fep\_indenture
7. fep\_new\_personnel\_or\_promotions
8. fep\_purchase\_agreement
9. fep\_remain\_as\_personnel
10. fep\_resignation\_or\_leaving
11. fep\_rights\_plan\_issue\_offer\_or\_agreement
12. fep\_securities\_purchase\_agreement
13. fep\_stock\_offering
14. fep\_stock\_option\_agreement\_or\_plan
15. fep\_stock\_or\_note\_conversion
16. fep\_stock\_purchase\_agreement
17. fep\_stock\_split
18. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
19. fep\_will\_or\_has\_generated\_income
20. fep\_will\_or\_has\_generated\_loss
21. fep\_will\_or\_has\_generated\_revenue

## **Appendix 24: Different FEP Types Recognised in the Second S&P 500 Dataset**

### Downs:

1. fep\_accountant\_dismissal
2. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
3. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill
4. fep\_amend\_rights\_plan\_issue\_offer\_or\_agreement
5. fep\_could\_would\_can\_expect\_material\_adverse\_effect
6. fep\_new\_personnel\_or\_promotions
7. fep\_purchase\_agreement
8. fep\_remain\_as\_personnel
9. fep\_resignation\_or\_leaving
10. fep\_rights\_plan\_issue\_offer\_or\_agreement
11. fep\_securities\_purchase\_agreement
12. fep\_stock\_or\_note\_conversion
13. fep\_stock\_purchase\_agreement
14. fep\_stock\_split
15. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
16. fep\_will\_or\_has\_generated\_income

### Ups:

1. fep\_accountant\_appointment
2. fep\_accountant\_dismissal
3. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
4. fep\_dividend\_distribution
5. fep\_indenture
6. fep\_meets\_or\_expects\_to\_meet\_listing\_requirements
7. fep\_new\_personnel\_or\_promotions
8. fep\_private\_placement
9. fep\_purchase\_agreement
10. fep\_remain\_as\_personnel
11. fep\_resignation\_or\_leaving
12. fep\_securities\_purchase\_agreement
13. fep\_stock\_or\_note\_conversion
14. fep\_stock\_purchase\_agreement
15. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
16. fep\_will\_or\_has\_generated\_income
17. fep\_will\_or\_has\_generated\_revenue

### Downs and Ups:

1. fep\_accountant\_appointment
2. fep\_accountant\_dismissal
3. fep\_acq\_ag\_and\_plan\_of\_merger\_or\_reorg
4. fep\_allocation\_writeoff\_or\_amortization\_of\_goodwill

5. fep\_amend\_rights\_plan\_issue\_offer\_or\_agreement
6. fep\_could\_would\_can\_expect\_material\_adverse\_effect
7. fep\_dividend\_distribution
8. fep\_indenture
9. fep\_meets\_or\_expects\_to\_meet\_listing\_requirements
10. fep\_new\_personnel\_or\_promotions
11. fep\_private\_placement
12. fep\_purchase\_agreement
13. fep\_remain\_as\_personnel
14. fep\_resignation\_or\_leaving
15. fep\_rights\_plan\_issue\_offer\_or\_agreement
16. fep\_securities\_purchase\_agreement
17. fep\_stock\_or\_note\_conversion
18. fep\_stock\_purchase\_agreement
19. fep\_stock\_split
20. fep\_to\_amend\_general\_ag\_report\_doc\_or\_info
21. fep\_will\_or\_has\_generated\_income
22. fep\_will\_or\_has\_generated\_revenue

## **Appendix 25: Different FEP Types Appearing in C4.5 Decision Trees**

- `fep_accountant_appointment`
- `fep_accountant_dismissal`
- `fep_acq_ag_and_plan_of_merger_or_reorg`
- `fep_allocation_writeoff_or_amortization_of_goodwill`
- `fep_dividend_distribution`
- `fep_new_personnel_or_promotions`
- `fep_purchase_agreement`
- `fep_remain_as_personnel`
- `fep_resignation_or_leaving`
- `fep_rights_plan_issue_offer_or_agreement`
- `fep_securities_purchase_agreement`
- `fep_stock_or_note_conversion`
- `fep_stock_purchase_agreement`
- `fep_to_amend_general_ag_report_doc_or_info`
- `fep_will_or_has_generated_income`