Sentiment Analysis for Financial News

Andrew Mayne

University of Sydney, 2010

Spike Shuffles After $13m Loss. Y2K Fears Trigger Spike In Trading Of 90 Day Paper. Sonic Back On Track, Beats Expectations. OFM pockets $1.3m profit on Little Lonsdale sale. News Corp drags med going for Newcrest. d. A senior member of ed calls for former Aris n Ducker to be denied t after he oversaw sh $2 billion. Multiplex Funds $3.6m Stadium Changes. Green Light For Mirvac's Magenta. Lend Lease's Foreign Hurdle. Investa makes $1.5bn POF bid. Harvey Norman Pushes US Expansion. Greens urge debate after media complaints upheld. Powers To Keep Fairfax Stake. Aluminum giant Alcoa could pocket as much as $120 million from the sale and leaseback of its Yennora rolling mill site in western Sydney. Loyal Customers Deserve A Rate Cut, Says Willis. CBA Problem Still Its Pricy Customers. Stanbroke Takes The Country Road. Caltex Expects Second Half Lift. CSR Managers Cash In On Price Rebound. ANZ Strategy Earns Analysts’ Crown. Computershare sticks to guns on profit. Cochlear Plans R&D Boost. Rivkin, Kennedy Accounts Frozen. Investors Have A Taste

Statement of Achievements

I have spent a year investigating techniques and methodologies involved in the fields of Finance, Natural Language Processing, Sentiment Analysis and Machine Learning. Combining these large fields of research, I have written an experimental system in the Python programming language that analyses financial news articles, and builds a statistical language model learnt from labelled data. The data is based on a novel corpus constructed with this task in mind, using 6,287 annotated articles sourced from the Australian Financial Review (AFR) over the period 1995-2006. Machine learning algorithms have been experimented with to build statistical models that are able to predict with significant accuracy and speed, the sentiment polarity of a news article (positive, negative or neutral), and its effect on an ASX listed company's share price. Using financial theory, I have examined how such a relation exists between news and stock prices, and exploited this using a novel trading algorithm and simulator. The positive trading returns achieved in various simulations, both confirm often debated financial theory and show the potential for such a system to be leveraged in real-world financial decision making and algorithmic trading strategies.

Abstract

Recent developments in the availability of machine-readable news has led to large numbers of researchers and practitioners to pursue methods in which to process and exploit such rich sources of data. These methods are not only a result of the explosion in the availability of information, but also come from the desire of how one can effectively extrapolate specific details from within text and combine these into a decision making process. This treatise investigates the use of Natural Language Processing (NLP) and Machine Learning methods to capture and predict underlying sentiment expressed within financial news. This treatise shows that by building statistical language models, one can predict the sentiment polarity of news articles and use this information in algorithmic trading scenarios to achieve significant positive return on the share market.

Acknowledgements

I would like to thank my parents, Peter and Louise, and my grandmother, who have always been there and supported me throughout school, university & and life; Belle for sending me hilarious emails and always making me jealous of her London tales; Eloise for steering me through the world of investment banking; James & Jackie for bringing some much needed entertainment to our home with Henry & Beatrice; Humphrey for greeting me every time as if I haven't been home for 6 months; The Vending Machine on SIT Level 1; Alcohol for helping me sleep after 7 coffees; Red Bull for being there when coffee just wasn't enough and my supervisors Rafael and Robert for their help and guidance for this project.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

The rapid expansion and proliferation of both qualitative and quantitative information in all forms of media has led many researchers to pursue methods to facilitate large-scale cultivation and manipulation of such data. The exponential growth of these knowledge resources has meant that it is near impossible for a human to read and process all the avenues available at their fingertips. This interesting and recent phenomenon has led to a great deal of research in the field of computational linguistics and information retrieval, into how one can use computational methods to refine, manage and exploit vast amounts of free text language resources. This treatise investigates the application of a variety of Statistical Natural Language Processing (SNLP) and affective methods to the financial news domain, and in particular, the semantic effect that a news article may have on stock market reactions. Research suggests that both the informational and affective aspects of financial news influence the market in profound ways, impacting on trade volumes, stock prices and volatility [1]. Trading algorithms that machine read news articles can react to breaking developments at lightning speeds, faster and more consistently than any human operators. This treatise will examine semantic analysis techniques in financial news, and evaluate their effectiveness in predicting movements in particular stock prices.

Newspaper articles, although seemingly objective, often express an author or commentator's opinion while reporting on recent events. Sentiment analysis, a type of subjectivity analysis [2], seeks to identify viewpoints underlying in text and recognize polarity or semantic orientation through analysis of words and phrases. The system presented aims at detecting references to given subject matter (such as price, income, resources) and leveraging contextual polarity information around the term and in the abstract as a whole. The abstract is used in place of the whole news article, as it generally summarizes the article's content in one or two sentences, contains rich affective content, and at the same time narrows the system's scope down to phrase and sentence level analysis. Document level sentiment analysis has proved to be a difficult task in the past, as it typically fails to detect subjectivity about individual aspects of a topic [3], [4], [5]. Confusion over varying topics and inflection is thus reduced by the use of abstracts, which typically only discuss one or two topics. The ability to predict stock market movements and behaviour is of much interest to researchers and critical to any market participant seeking positive returns. Behavioural patterns and investor sentiment are not only complex phenomena, but are inherently subjective. Sentiment in the area of finance, as in all other linguistic domains, cannot be wholly defined or comprehended, and is subject to constant change, thus making accurate predictions rather difficult [6]. This leads to the questions: *What makes a news article change the reader's view of a company? What informative content persuades the reader to make such a decision? Does an investor make this decision on this one article, or on previous research?* Of course, answers to these questions cannot be fully measured or even extrapolated quantitatively, but research certainly suggests market participants are heavily influenced by informational releases [7].

While some sources of information, such as blogs or forums, can have material effect on market opinions, news articles are considered a more stable and commoditised form of information [8]. Not only does this news affect the primary equity and fixed income markets, but also their respective derivative markets. This type of analysis is currently being exploited by commercial trading applications to help manage event risk and generate alpha returns, such as Reuters Newsscope Sentiment Analysis [9]. With the advent of cheaper processing and knowledge acquisition techniques, the goal of such applications is to apply automated versions of existing fundamental and technical strategies, and effectively remove the elements of emotions and biases from trading by humans [10]. To further create confusion, there are two diametrically opposed philosophies of financial market research; fundamental and technical analysis [11].

CHAPTER 2

Background

2.1 Financial Background

2.1.1 Efficient Market Hypothesis

For at least 45 years, financial economists have formally considered the notion of informationally efficient markets, and the importance of the Efficient Market Hypothesis (EMH). An 'efficient' market is defined by Fama in [12], as a market where there are large numbers of rational, profit-maximisers actively competing, with each trying to predict future market values of individual securities, and where important current information is almost freely available to all participants. The EMH asserts that is impossible to consistently achieve positive returns in financial markets because existing share prices always incorporate and reflect all relevant information. As a consequence of EMH, the only way an investor can outperform the market is not through expert stock selection or market timing, but through the purchasing of riskier investments that yield a higher return. Most evidence from studies in finance show that markets are at least weak-form efficient, and generally semi-strong form efficient [13].

Weak form efficiency insists that stocks are exceptionally tricky to value, and investors may find it easier to price a company's shares relative to yesterday's price, or relative to another securities'. In other words, they generally take yesterday's price as correct, adjusting upwards or downwards on the basis of today's information. However, when investors lose confidence in yesterday's price, there may be a period of confused trading and volatility before a new benchmark is established. In a semi-strong form efficient market, "you can trust prices as they impound all publicly available information about the value of the security" [14]. Supporters of the EMH believe it is pointless to search for undervalued stocks or perform fundamental or technical analysis. However, while a large body of academics point towards evidence in favour of EMH, equal amounts of dissension also exist, as it is a highly controversial and often disputed theory. Interestingly, this opposition rears its head particularly in severe market downturns, such as the 1987 stock market crash, or the more recent Global Financial Crisis (GFC) in 2008. Brealey notes in [14] "the EMH emphasises that arbitrage will rapidly eliminate any profit opportunities and drive market prices back to fair value" but also argues that individual investors have built-in biases and misperceptions that can push prices away from fundamental values. These biases and misperceptions are typically driven by market sentiment through media outlets such as news, blogs, forums and colleagues. It is this phenomenon that this paper wishes to exploit. Evidence of longer-term inefficiencies such as the apparent long-term delay in the reaction to news [15]. Through this large body of research conducted on informationally efficient markets there seems little left to learn from continued empirical examinations of information and markets [7].

2.1.2 Financial Sentiment

From the recent disasters in financial markets, it may seem that investors are affected by a bounded rationality [16], through a self-deception in rejecting any new evidence about stocks in favour of prior (and possibly incorrect) information, such as that seen in the ".com" bubble [17]. An investor's sentiment is often derived from a variety of news and data sources, often by relying on some and discounting others in certain situations [18].

Although this task is fundamentally very subjective, the notion of sentiment within the financial domain is slightly simplified in comparison with other domains, due to the causal relationships between key indicators (such as net income, tax, sales etc) and a company. Sentiment is drawn from 'good prospects' and future profits for a company, not just attitude expressed by opinion holders. For example, tax is generally seen as a neutral topic linguistically, however a rise in taxes or regulation could be very detrimental to an investor's profit and it may be the case that the news writer is objective in their reporting on this event. These relationships are not necessarily captured by traditional linguistic sentiment, but can have a great impact on the sentiment of a market participant towards stock(s). Thus, this paper reasons that sentiment in the financial domain is not only the linguistic sentiment conveyed by the author, but also the sentiment implied by changes in influential economic and financial indicators. The system presented attempts to address this causality relationship using a novel approach.

It is important to note that this paper does not try to affirm or falsify any financial theory, but merely use it to suggest a series of constructs (hypotheses) on which the system should be based and evaluated.

2.2 Natural Language Processing

In contrast to all the technical developments that the web has facilitated, human capabilities to process this influx of information available have not increased likewise. In analysing any language, in particular modern English, one must be aware of the inherent complexities and constant changes that a language is subjected to over time. As natural languages are essentially fluid, dynamic and rich phenomenon, linguistic models very easily become convoluted, with exceptions to the rule quickly emerging and becoming apparent. Herein lies the difficulty in computational linguistics: *What does a word mean to a computer?* Their representation in memory is merely a sequence of characters in a string. In many modern languages, words are defined through a complex, deep set of features, including: orthographics, semantics, phonology, morphology, grammar and syntactics [19]. Words can convey meaning, moods and opinions. Machines must learn and use mathematical models to comprehend language, through probability theory and a variety of statistical measures, to predict relationships typically extrapolated on previously human-annotated corpora. These corpora are not cheap or easy to come by, as they typically involve large amounts of time and manual processes by a domain expert (such as a linguist).

2.3 Sentiment Analysis

Although humans can easily recognize opinions in text, this process is not easily translated into machine algorithms with high degrees of accuracy. The analysis of favourable and unfavourable opinions is a task that requires high intelligence and deep understanding of the textual context, drawing on common sense, domain and linguistic knowledge [20], with full comprehension remaining well beyond the power of machines. However, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts entities [21]. The analysis of large amounts of textual content is of huge importance to many people and organisations. Corporations, for example, seek to find consumer and public opinions on their particular products and services, and likewise, potential consumers seek experiences of existing users before making any consumption decisions [22]. Sentiment analysis can also facilitate policy makers or politicians to analysis public sentiment in regards to policy and political issues [23]. Much of the existing work in the field of sentiment analysis is based upon product or movie reviews, such as [3], [22], [24] & [25]. One reason for this is the fact that these reviews typically link textual reviews with an overall scoring system (*"Thumbs up, thumbs down"* approaches), and thus provided easily accessible annotations to the derived corpora [26].

The first sub-task in sentiment analysis is typically document classification, with which each document relates to particular class (such as product, movie, book etc). Class disciplines can be of varying degrees of granularity, specific to the task, but remain important to distinguish which events correlate with which opinions [27]. This paper refrains from extensively investigating any document classification techniques, as the corpus provides the system with any class information necessary. Due to the nature of companies, a certain stock is typically related with a specific industry (specifically, a Global Industry Classification Standard (GICS)). See Table 2.1. The system leverages this information, gaining insight to domain-specific terms, such as *'iron ore'*, *'aluminium'* or *'optic fibre cables'*. These terms will provide us with important cues for event identification in respect to an opinion holder, as well as causal relationships between directional aspects (i.e. *up* and *down*, *rise* and *fall*) and key words (*sales*, *resource X*, *technology Y*). A break-down of news topics can be found in Figure 2.1. In [28], Esuli and Sebastiani identify 3 specific subtasks that make up sentiment analysis; subjectivity, polarity and polarity strength, discussed in the following sub-sections.

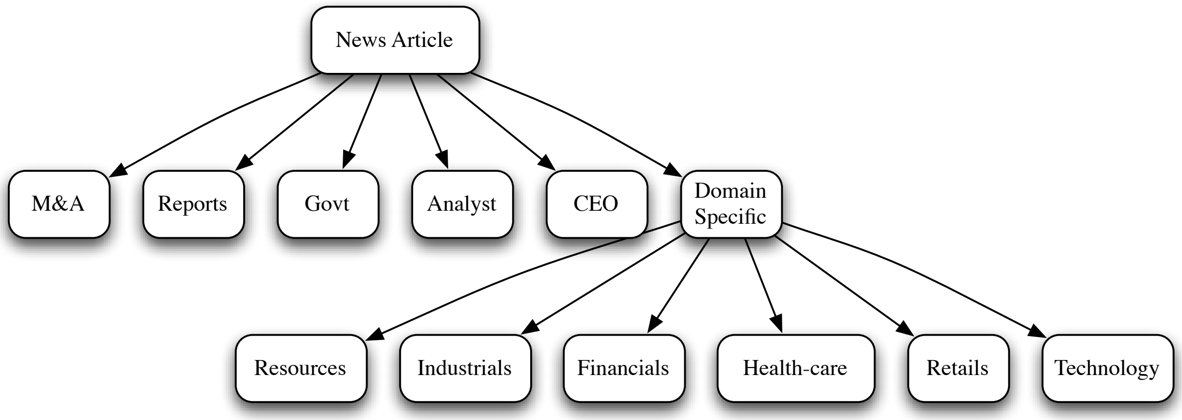

FIGURE 2.1: A breakdown of news topics or 'events'

2.3.1 Subjectivity

Determining subjectivity involves deciding whether a piece of text is factual or subjective. Subjective classification decides whether sentences in each abstract carries opinion, on the words and structure used by the author. Subjectivity may be detected by the aggregation of sentimental features (such as adjectives) within a sentence. However sentences, such as that seen in Pang and Lee [29], may carry sentiment without any (computationally) obvious clues at all: *"If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut"*. This sentence shows how hard it can become in deciding whether a text contains objective fact or expresses judgment (in this case, sarcasm).

2.3.2 Polarity

Managing sentimental polarity involves deciding whether, given a opinionated sentence, it carries either a positive or negative viewpoint. Opinion mining is a recent subdiscipline at the cross roads of information retrieval and computation linguistics, which aims at determining the subjectivity expressed within a document. SentiWordNet provides a high-coverage lexical resource which will be leveraged heavily by the system presented [30]. SentiWordNet is based upon the quantitative analysis of the glosses of a semi-supervised synset classification of the WordNet resource, assigning each synset a triplet of scores $\Phi(s, p)$ (for $p \in P =$ $\{$Positive, Negative, Objective$\}$).

2.3.3 Polarity Strength

Polarity strength is important in this domain as it indicates the intensity of opinion, which could potentially translate to the confidence of the author in this event/topic. SentiWordNet provides a quantitative strength indicative of how positive or negative a synset may be, however this may not be a strong enough resource in this domain. Other resources such as the financial gazetteer may help in this case, strengthening the author's opinion through directional features (up or down, and by how much).

2.4 Machine Learning

Machine learning is a scientific discipline that allows one to extrapolate behaviours based on empirically gathered data. Their ability to learn complex behavioural patterns and make intelligent decisions make them a perfect tool for such elaborate phenomenon as linguistics. Sentiment analysis is essentially a classification task, in which categories (negative, positive, neutral) must be assigned to data sets with given feature vectors. The features used in opinion analysis

No.	GICS Industry Sector & Industrial Groupings
1	*GICS Australian Real Estate Investment Trusts*
2	*GICS Consumer Discretionary* Consumer Services Automobile & Components Media Retailing Consumer Durables & Apparel
3	*GICS Consumer Staples* Food Beverage & Tobacco Food & Staples Retailing Household & Personal Products
4	*GICS Energy*
5	*GICS Financials* Diversified Financials Banks Insurance
6	*GICS Health Care* Health Care Equipment & Services Pharmaceuticals, Biotechnology & Life Sciences
7	*GICS Industrials* Transportation Capital Goods
8	*GICS Information Technology* Semiconductors & Semiconductor Equipment Software & Services Technology Hardware & Equipment
9	*GICS Materials* Materials Metal & Mining
10	*GICS Telecommunication Services* Utilities Telecommunication Services

TABLE 2.1: List of GICS Industry Sector & Industrial Groupings

are of a much more complex nature than text categorisation tasks, as they are tied to linguistic subjectivity and not simply terms observed within a text. As such, the task requires powerful machine learning techniques to learn detailed behavioural patterns on extremely sparse data sets.

CHAPTER 3

Design

3.1 Corpora Data

The corpus used in this treatise has been sourced from *The Australian Financial Review*, Australia's leading financial newspaper, published for more than 50 years. The AFR is Australia's highest reaching publication amongst senior business executives and has a weekly readership of 243,000 [31], and it is hoped that news reported here will have material affect on a company's stock price. The corpus contains all articles relevant to particular companies in Australia and listed on the ASX, for the years 1995-2006. For each article, only the headline and the first paragraph are retrieved, as these typically contain the most significant news. Perhaps more importantly, analysing documents of greater length have proven to be a difficult task in the past, with both sentiment and topic changes throughout (such as product and movie reviews in [22]), and for this reason the rest of the article is discarded.

3.1.1 Annotation

Corpus annotation is an important first task before any experimental design is undertaken, as it provides not only a gold standard from which to learn, but also with which to evaluate upon. In generating one's own corpus, there are several design issues that need to be considered, in

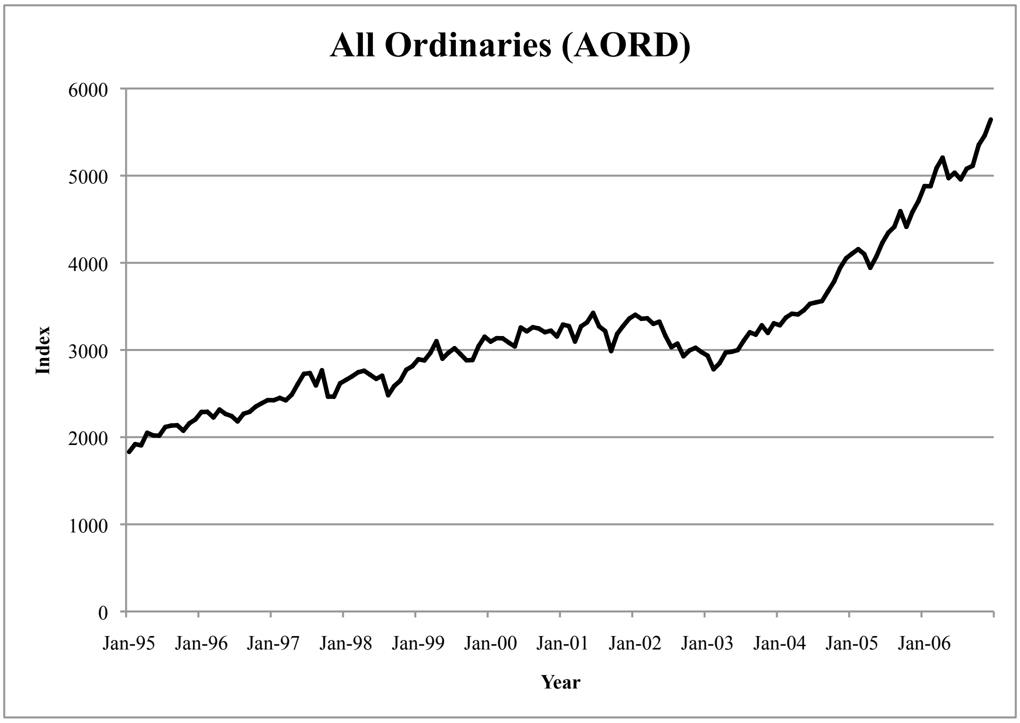

FIGURE 3.1: The All Ordinaries Index over the period of 1995 to 2006, in which one can see a sustained rise in the index over time, accounting for a large number of positive articles in the corpus

order to create a corpus that is representative of the task at hand. By sampling over 11 years of news articles, it is hoped that statistical noise is reduced and an even distribution of annotation classes can be obtained. However due to the nature of financial markets over the period 1995-2006, as seen in Figure 3.1, there is a natural skew towards positive news. Of the articles annotated, 45% were positive, 27% were negative and 28% were neutral. The breakdown of the corpus by GICS Sector Industries can be found in Figure 3.2.

Annotating ones own corpus is also the most costly venture involved in a project such as this, as it requires time investment from annotators with background knowledge of the subject (Annotation productivity ranged from 100-200/articles per hour). Previous papers on financial sentiment analysis, such as [32] have reported very low agreements scores, with Kappa as low as 54%. Obtaining double annotations for all articles in the corpus over the course of this treatise would be out of the scope of any budget, so a small subset of articles were chosen to be annotated by two independent annotators. 6,287 articles were annotated by single annotators, and of those, 1,665 were annotated by two annotators, resulting in a promising Kappa of 70.64%

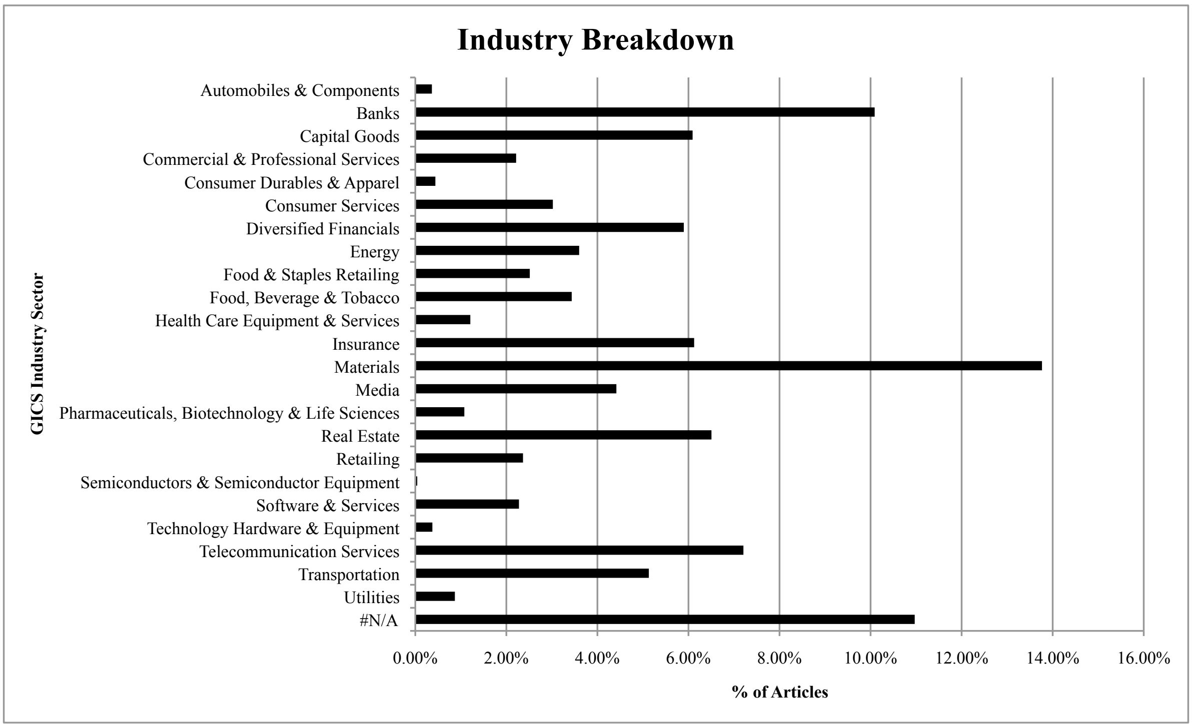

FIGURE 3.2: This figure shows the percentage of articles within each GICS Industry Sector

3.1.1.1 Annotation Guidelines

Annotation Instructions for Sentiment Analysis Experiment

This document provides instructions for carrying out data annotation for a Sentiment Analysis experiment, in which we want to determine whether software can be developed to mimic people's perceptions of the sentiment expressed in newspaper articles. In order to do this, we have accumulated a large amount of textual data, which requires annotations by human subjects.

Introduction
You will be required to examine abstracts from newspaper stories relating to particular companies, sourced from the Australian Financial Review. No prior knowledge of the company is required, simply a background in practical financial concepts. We want you to examine each abstract in terms of the positive or negative nature of sentiment expressed towards the company. This is not only the author's sentiment towards the subject matter, but also the nature of the subject matter itself. In other words, whether you believe this article will have a positive or negative impact on the company itself.

The data is presented in the form of an Excel spreadsheet (CSV file), with each row representing a particular news article.

For each article, you'll see the following columns:
- Document ID – this is for data purposes only, and you can ignore this
- Ticker – The company's ASX ticker code
- Company Name – The company's full name
- Company Industry – The GICS Industry Sector the company belongs to
- Article Headline – The headline of the article
- First Paragraph – The first paragraph of the article
- Sentiment – The sentiment score of the article, which you will enter, using one of the three categories defined below

Sentiment Polarity
Consider whether the article's author expresses a positive, neutral or negative sentiment towards the company, and whether you believe that this will have a positive, negative, or no impact upon the value of the company. It is also important to view this impact in isolation, and not to draw from any contextual knowledge you might know about the particular company (ie. its current prospects, or the price at the time etc.). Please annotate each abstract with the following annotations:
- *Positive* - This news will have a positive impact on the company
- *Negative* - This news will have a negative impact on the company
- *Neutral* - The impact of this news on the company is ambiguous, or no material impact

Examples
The following are some example annotations with which you can use as a guideline:

Abstract	**Annotation**
Company X experienced a large rise in net income for this quarter	Positive
Company Y continues strongly with its expanding operations in Tasmania	Positive
Company Z set sights on new mining project worth $500 million	Positive
Company X has reported weak profit results for fourth quarter, far below analysts' expectations	Negative
Company Y abandoned merger talks in light of a weakening Australian Dollar	Negative
Company Z faces regulatory fight over new project	Negative
Company X's recently purchased new works of art for its city office	Neutral
A new managing directory for Company Y	Neutral
Mr. Smith takes Company Z's top job	Neutral

We thank you in advance for your assistance in this experiment.

3.2 Feature Extraction

Feature extraction is an integral part of this system design, in order to prepare the articles for feature selection and subsequently machine learning. In this pre-processing step, each article is parsed and tokenised for use as input into the C&C Parser, which performs Part of Speech (POS) Tagging and Named Entity Recognition.

3.2.1 Tokenisation

Tokenisation is the process of breaking up strings of words into smaller, more manageable 'tokens', enabling the system to segment text into lists of tokens for further processing. At the document level, sentence tokenisation occurs first with word tokenisation following. The C&C Parser, used in the feature extraction stage, requires sentence-level tokenisation as input. At first glance, this can be seen as a simple task, simply splitting a document at each full stop '.', however this does not handle any abbreviations, such as *'U.S.A.'*. To enable proper sentence tokenisation, regular expressions were utilized to properly split each document into sentences for input into the C&C Parser.

3.2.2 C&C Parser

The C&C Parser, written by Curran & Clark [33], is based upon Combinatory Categorial Grammar, and helps with the first step in extracting features from the text. It is written in C++ and is available as a SOAP service, so is efficient enough for large scale language processing tasks that this has the potential to be. It provides accurate tokenisation, Part of Speech tagging and Named Entity Recognition for preparing an article for feature selection techniques.

3.2.2.1 Part of Speech

The English language has eight main parts of speech or lexical classes: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. Parts of speech do not explain a word itself, but the functional role a word form may use within a sentence and are generally defined by either the syntactic or morphological behaviour of the word in question. In linguistics, one makes a further distinction between lexical classes, by open and closed word classes. Open word classes are defined as a lexical class that accepts the addition of new items through compounding, borrowing etc. (and are typically nouns, verbs and adjectives). Closed word classes, on the other hand, are a closed set of words that generally remain unchanged, which include prepositions, postpositions, determiners, conjunctions and pronouns. For more information see [19] and [34]. This system presented incorporates part of speech taggers trained on the Penn TreeBank corpus. The Penn Treebank corpus was the first large-scale treebank to be published, which includes 2499 stories from the Wall Street Journal in 1989, and is annotated both with part of speech and syntactic tree structure (i.e. a Treebank). For a complete set of part of speech tags in the Penn Treebank, refer to Table 3.1. To perform accurate part of speech tagging, Hidden Markov Models (HMMs) are implemented, which essentially disambiguate speech based on contextual words, and probabilities of sequences observed in the training data, and achieve accuracies of 93-95% [35], [36]. Charniak however notes that in the Penn Treebank, merely assigning the most common tag to each known word and tag unseen words as proper nouns, one will approach 90% accuracy, as many words in the corpus are unambiguous [37]. For an example of part of speech tagging output, see Figure 3.3.

AMP	leads	Asian	markets
NNP	*VBZ*	*JJ*	*NNS*
Proper Noun	Verb 3rd ps. present (singular)	Adjective	Noun (singular)

FIGURE 3.3: Part of Speech Tagging Example

No.	Tag	Part of Speech	No.	Tag	Part of Speech
1	CC	Coordinating conjunction	25	TO	To
2	CD	Cardinal number	26	UH	Interjection
3	DT	Determiner	27	VB	Verb, base form
4	EX	Existential there	28	VBD	Verb, past tense
5	FW	Foreign word	29	VBG	Verb, gerund or pres. participle
6	IN	Preposition or subord. conjunction	30	VBN	Verb, past participle
7	JJ	Adjective	31	VBP	Verb, non-3rd ps. sing. present
8	JJR	Adjective, comparative	32	VBZ	Verb, 3rd ps. sing. present
9	JJS	Adjective, superlative	33	WDT	Wh-determiner
10	LS	List item marker	34	WP	Wh-pronoun
11	MD	Modal	35	WP$	Possessive wh-pronoun
12	NN	Noun, singular or mass	36	WRB	Wh-adverb
13	NNS	Noun, plural	37	#	Pound sign
14	NP	Proper noun, singular	38	$	Dollar sign
15	NPS	Proper noun, plural	39	.	Sentence-final punctuation
16	PDT	Predeterminer	40	,	Comma
17	POS	Possessive ending	41	:	Colon, Semi-colon
18	PP	Personal pronoun	42	(	Left bracket character
19	PP$	Possessive pronoun	43	)	Right bracket character
20	RB	Adverb	44	“	Straight double quote
21	RBR	Adverb, comparative	45	‘	Left open single quote
22	RBS	Adverb, superlative	46	“	Left open double quote
23	RP	Particle	47	’	Right close single quote
24	SYM	Symbol	48	”	Right close double quote

TABLE 3.1: List of all Penn Treebank Part of Speech Tags

3.2.2.2 Combinatory Categorial Grammar

Categorial Grammar (CG) is one of the oldest lexicalised grammar formalisms [38] [39] and has made significant contributions towards the study of semantics, syntax, morphology, international phonology, computational linguistics and human sentence processing [40]. In CG, all grammatical constituents are distinguished by syntactic type, and generally combine as functions or according to a function-argument relationship. Such types, or categories, are transparently related to the semantic type of the linguistic expression itself, and specify language-specific linear order. Pure categorial grammars include only forward and backward functional application combinators. Indeed, one can see that these grammatical categories and their formalism in expression, are transferable to the realm of machine translation [41]. More recent statistical parsers have used models based on lexical dependencies [42], typically derived from a context-free phrase structure tree with simple head percolation heuristics. These heuristic rules are recovered from latent information within the Treebank corpus [43]. However these approaches tend not to work well for longer-range grammatical dependences involved in type-raising, control, extraction and coordination, all of which are common in text such as the Wall Street Journal [44]. Clark et al. present a novel approach to explore deep semantic dependency structures within language using the Combinatory Categorial Grammar (CCG), the wide coverage C&C parser and Boxer in [33]. There have been several previous attempts at making use of semantic information in text, such as logical representations in [45], however such work is yet to be embraced by the community at large [46]. In [45], Bos and Nissim make use of the CCG parser to interpret and identify superlatives within sentences, and these can then be combined with information extraction techniques to identify objects and their features. CCG formalism defines three primitive types of lexical items, *S*, *N* and *NP* as well as more complex expressions, such as *S\NP* or *NP/N*.

3.2.2.3 Named Entity Recognition

Named Entity Recognition (NER) is a subtask of information extraction, in which the goal is to seek and classify entities within the text to be used as features in the machine learning step. This is particularly important when using contextual windows, to encapsulate subjective words surrounding an entity, such as a company, for instance *"fantastic Commonwealth Bank profit results"*. C&C Tools performs this with state-of-the-art accuracy, with a low granularity based on person, location and organisation, which is perfect for the domain of finance.

3.2.3 Lemmatisation

Zipf's law states that given a corpus of natural language, the frequency of any word is inversely proportional to its rank in a frequency table (see Figure 3.4). Put simply, the most frequently occurring word will occur twice as often as the second most frequently used word. Due to the natural Zipfian distribution of natural language [Manning & Shutze], lemmatisation is employed to the corpus to provide a more informative statistical description of the text. Where stemming refers to the stripping of inflectional affixes from words, lemmatisation refers to the algorithmic process of determining the lemma or root of a word from its inflected form and part of speech. While stemming the words 'walks','walking' and 'walked' to 'walk' is straightforward, other inflected forms such as superlatives add a greater degree of complexity. For example, 'good', 'better' and 'best' share the lemma 'good', which is missed by traditional stemmers. Lemmatising words within a corpora theoretically reduces the number of features, and hence the sparsity of the data, providing a smoother, more revealing distribution. In this pre-processing step, WordNet's Lemmatiser is used, which is included in the NLTK code package.

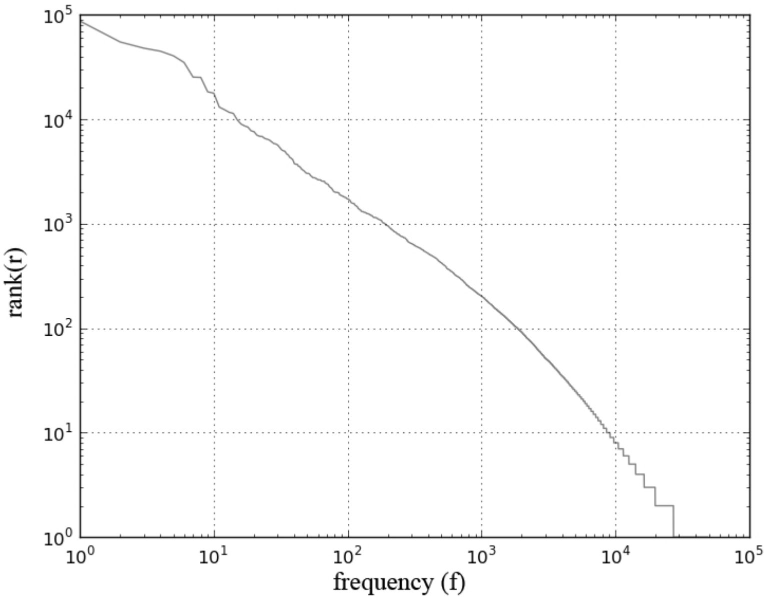

FIGURE 3.4: This plot (Rank versus Frequency), exhibits the Zipfian (Hyperbolic) distribution of words, using logarithmic scales, generated from the AFR corpus. This shows that the frequency of any word is inversely proportional to its rank.

3.2.4 Ngram Words

As part of pre-processing, tokenisation splits the raw textual data into small, meaningful word units, commonly referred to unigrams. Unigrams are a typical baseline for many language processing experiments (a 'bag-of-words' approach), and provide a measurement for the system to outperform. However, often words do not only occur as single logical units. Often separating on white-space alone is not enough to encapsulate the meaning of some words, such as 'net income' and 'earnings per share'. Manning and Shutze in [47] define collocations, or Ngram words, as being characterized by very limited compositionality; an expression is compositional if the meaning of the whole expression can be predicted from the meaning of the parts. For a system to be effective in capturing such a complex meaning as sentiment, bigram and trigram words must be included in order to effectively process language in the text.

3.2.5 Boolean Weighting

Although a simplistic approach, Boolean weighting is found to be extremely powerful in many classification tasks [48], particularly in cases were the frequency of a word in each document is likely not to be more than one (which is the case in the AFR corpus). For each feature in the vector, weighting will be assigned either 1 or 0 to indicate presence within a document:

$$x_{ik} = \begin{cases} 1 & \text{if } a_{ik} \text{ is present} \\ 0 & \text{otherwise} \end{cases}$$

3.3 Feature Selection

Feature Selection is an important task in document classification as it enables us to reduce the dimensionality of the data set without losing any of its descriptive nature. A subset of features is selected in order to build more robust learning models, with several statistical techniques tested to create the most descriptive feature set for a machine learning algorithm to train the model, including Term Frequency, Inverse Document Frequency and Entropy.

3.3.1 Stopwords

Because of the Zipfian phenomena, the distribution of words in a corpus are naturally skewed by frequently occurring, non-informative words that provide the ‘glue’ between text, such as ‘a’, ‘an’, ‘be’, ‘do’, ‘from’, ‘the’, ‘where’ etc. These are commonly referred to as ‘stop words’ and include non-informative articles, prepositions and other function words. Such words are useful in pre-processing steps, through POS Tagging and Named Entity Recognition, however tend to attribute large amounts of statistical mass to non-informative words, and hence must be filtered out. A list of stop-words is provided by the NLTK package, and are removed from the feature subspace for this classification task.

3.3.2 Term Frequency

Although a simple approach, counting the frequency of each word and then thresh-holding based on rank is often the first technique applied to document classification experiments, producing good performance gains. Each word’s frequency is counted over the whole corpus, put into a ranking table, and the top 100 is chosen. This process is applied after stop words and punctuation have been removed, and is performed for unigram, bigram and trigram terms using NLTK’s inbuilt collocation finder.

3.3.3 Inverse Document Frequency

Inverse Document Frequency (IDF) is another common approach in Language Processing experiments [49], which takes into the occurrence of a word throughout all documents in a corpus. Instead of simply counting the frequency of each word, IDF weights each word in inverse proportion to the number of the documents in the collection which have such a word. A word that is more unique to one type of document as opposed to another type will thus carry more importance in classification and more strongly discriminate between articles in the corpus. The

IDF is calculated based on the industry that each company is in, primarily to capture industry specific words described in Figure 3.2. Typically TFxIDF is used, where the term frequency is multiplied by its inverse document frequency, however, since the corpus contains short articles of typically 50 words, it is unlikely that a word will occur more than once. As such, simple IDF is threshholded over the average IDF score if a word appears in a document like so:

$$x_{ik} = \begin{cases} 1 & \text{if } \log(\frac{N}{n_i}) > \text{idf-average} \\ 0 & \text{otherwise} \end{cases}$$

3.3.4 Entropy

Entropy is a measure of the uncertainty associated with a random variable which, in the case of information retrieval, is how 'informative' a feature is pertaining to a certain class. According to Shannon, who first published his famous paper introducing the theory of Entropy in 1948, it is the amount of information content one is missing when one does not know the value of such a random variable [50]. Entropy is calculated for each word in the corpus, with respect to two categories, industry and company. Features are selected based on which word carries the highest (or maximum) entropy for each company and industry, which is based on the postulate that some words in an article may have higher significance based on which company/industry is being talked about (such as *'aluminium'* for a mining based company). The formulae for entropy is expressed quantitatively as the following:

$$H(X) = \arg\max[-\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)]$$

3.3.5 Gazetteers

Lexicons or Gazetteers are often used in NLP tasks as they represent domain specific words that are important in classification. Gazetteers are not always ideal due to their fixed nature, however their use is motivated by Finance and Accounting terminology, which is not set to change, such as ‘net income’ and ‘profit’. Two dictionaries were curated from the Internet, for typical Accounting and Finance terms, using VentureLine’s Accounting Dictionary [51] and Campbell R. Harvey’s Finance Glossary [52]. In addition, a manually constructed ‘directional’ gazetteer was used, motivated by tendency to use ‘directional’ words in reference to key financial indicators. For example, *“net income climbed by 42 per cent”* and *“recent projects were damaged by rising oil prices”*. This dictionary was aimed at finding words that are typical within financial news articles, such as ‘rise’ and ‘fall’. Using these as the two starting seeds, the dictionary was constructed from all words relating to direction in WordNet.

3.3.6 SentiWordNet

SentiWordNet is a lexical resource devised explicitly for supporting sentiment classification and opinion mining applications. SentiWordNet is the result of the the automatic annotation of all the synsets in WordNet according to the notions of “positivity”, “negativity” and “neutrality”, with each synset assigned three numerical scores, which indicate the sentiment polarity of the associated synset [30]. Each of the three scores range in the interval [0.0,1.0], and are related to their part of speech (reduced to Noun, Verb and Adjective/Adverb). If a word or it’s lemma has a non-zero score in either positivity or negativity (or both), it is included as a feature.

3.3.7 Contextual Features

Contextual windows are used as a feature selection step, which capture words in a neighbouring range of [-1,1]. If a token has been labelled as important (i.e. it has been found in a gazetteer etc.), the system includes words to the left and right as features that also have labels, in the form *'left token - term'* and *'term - right token'*. Importantly, this covers phrases of the form [adjective financial-term], such as *'rising profit'* and *'falling EPS'*.

3.4 Machine Learning

Statistical modelling addresses the problem of modelling the behaviour of random processes, in this case, natural language. Machine learning algorithms help researchers construct models based on empirical data, which succinctly represents such a (random) process based on partial knowledge. This compact representation in the form of a statistical model then helps in predicting the future behaviour of this process. The learner takes advantage of the fact that there is some unknown, underlying probability distribution in the data, automatically extracting the relationships from the data set and combining these rules into a predictive model. How the sample data is presented to the machine classifier is a crucial step in any information retrieval task, and is discussed in sections 3.2 & 3.3. In this treatise, three popular machine learners are used, Naïve Bayes, Maximum Entropy Modeling and Support Vector Machines, which have proved fruitful in many other NLP tasks (for example [6], [29], [53]). Each document (or article) is represented as a vector of features that have been extracted from the text. Each machine classifier constructs from this training data estimates of the probabilty of each class given the document feature values. Once a model is constructed, it can be used to make a class prediction based on the features present in each article.

3.4.1 Naïve Bayes

Much statistical learning stems from the partial knowledge of the outcome of an experiment and the conditional probability that the given evidence (x_i) will result in the hypothesis (c_j). Bayes' Theorem shows the relationship between conditional probability and it's reverse form [54], expressing the posterior probability $P(c_j|x_i)$ in terms of the prior probabilities $P(x_i|c_j)$, $P(c_j)$ and $P(x_i)$, generalising to:

$$P(c_j|x_1, x_2, x_3, ..., x_k) = \frac{P(c_j)P(x_1, x_2, x_3, ..., x_k|c_j)}{P(x_1, x_2, x_3, ..., x_k)}$$

The denominator $P(x_1, x_2, x_3, ..., x_k)$ does not differ across classes, so it is effectively constant and left out in calculation. The numerator is equivalent to the joint probability, denoted by $P(c_j, x_1, x_2, x_3, ..., x_k)$, which can be expanded by repeated applications of the definition of conditional probability as follows:

$$\begin{aligned} P(c_j, x_1, x_2, x_3, ..., x_k) &= P(c_j)P(x_1, x_2, x_3, ..., x_k) \\ &= P(c_j)P(x_1|c_j)P(x_2, x_3, ..., x_k) \\ &= P(c_j)P(x_1|c_j)P(x_2|c_j, x_1)P(x_3, ..., x_k) \\ &= P(c_j)P(x_1|c_j)P(x_2|c_j, x_1)...P(x_k|c_j, x_1, x_2, ..., x_{k-1}) \end{aligned}$$

This leads to high computational complexity, and as such the Naïve Bayes assumption eases computation, by assuming conditional independence between feature x_i and every other feature

$x_j (j \neq i)$, generating the joint probability model:

$$\begin{aligned} P(c_j|x_1, x_2, x_3, ..., x_k) &= P(c_j)p(x_1|c_j)p(x_2|c_j)p(x_3|c_j)...p(x_k|c_j) \\ &= P(c_j) \prod_{i=1}^{k} P(x_i|c_j) \end{aligned}$$

The Naïve Bayes classifier uses this joint probability model and combines it with a decision rule, namely that of maximum a posterior rule, which chooses the hypothesis which is most probable.

Naïve Bayes is often criticised because of its apparent unreasonable efficacy, through its 'naïve' assumption of conditional independence, as there are few real world problems that do not have varying degrees of dependency between features, and language processing is no exception. Through the independence assumption, Naïve Bayes can effectively 'double-count' highly correlated features. Although a trivial consequence of the definition of conditional probability and Bayesian belief (in which one can calculate $P(c_j|x_i)$ in terms of $P(x_i|c_j)$), and despite its conditional independence assumption, a Naïve Bayes classifier is surprisingly effective, even with small amounts of training data. And unlike the other two classifiers, Naïve Bayes naturally lends itself to multi-class classification, without any further modification. In this treatise, the NLTK's Naïve Bayes classifier is used, with a maximum a posterior rule, and is found to be much more efficient than comparable packages such as WEKA tools [55], due to such a large feature set coupled with the overhead of WEKA's implementation.

3.4.2 Maximum Entropy

The principle of Maximum Entropy is a natural extension to Bayesian theory (without Naïve Bayes' independence assumption), and is a probability distribution estimation technique widely used for a variety of natural language tasks, including language modelling, part of speech tagging and document classification [56]. The underlying assumption to maximum entropy as described in Section 3.3.4 is that without external knowledge, one should prefer the distribution that is most uniform [57]. Constraints on the distribution are learnt from a set of labelled training data, minimally shifting the distribution to non-uniform, and documents are classified by estimating the conditional distribution of the class variable. For example, if one had 3 classes and an unseen feature set, maximum entropy would guess a uniform distribution for all 3 classes - 33%. In its most general formulation, maximum entropy can be used to estimate any probability, and as one adds constraints through labelled data, a conditional distribution is formed. It is out of the scope of this paper to show the complete formulation of Maximum Entropy, as this could entail another paper in itself, thus a more complete proof can be found in [57] or [58]. In this treatise, the MegaM tool is used [59], which uses maximum likelihood and maximum a posterior optimisation in parametric form, a more efficient implementation as opposed to the more widely used iterative scaling technique.

3.4.3 Support Vector Machines

Support Vector Machines (SVMs) are a popular machine learning approach presented by Vapnik in [60], and has proved itself an efficient and accurate supervised machine learning technique. SVMs differ significantly in their methodology when compared to Naïve Bayes and

Maximum Entropy Modelling, by performing classification through constructing an N-dimensional hyperplane that optimally separates the data into two categories (typically -1,1). The goal of the SVM is to find an optimal division that separates clusters of vectors or data points. SVMs are a close cousin to the classical multilayer perceptron neural networks (NNs). However it holds a significant advantage over neural networks in that, whilst NNs can suffer from multiple local minima, the solution to an SVM is global and unique. And, unlike NNs, its computational complexity is not dependent on the dimensionality of the input space. Its ability to handle large feature sets and sparse data (in effectively linear training time) makes it an attractive machine learner for such complex tasks as opinion analysis. From a given set of feature vectors, SVMs deduce linear combinations of features from appropriate examples called support vectors. These support vectors define a hyperplane in this multidimensional feature space, separating (ideally all) positive examples from all negative examples, such as that seen in Figure 3.5. Joachims argues that SVMs offer two important advantages in the field of text categorisation: feature selection is often not needed, as SVMs tend to be fairly robust to overfitting and can scale up to considerably high dimensionalities in sparse data sets, and they require little parameter tuning due to default choice of parameter settings with best effectiveness [53].

3.4.3.1 The Linearly Separable Case

The principle of SVM is most easily understood via the case of linearly separable data. Data is linearly separable if there exists a line that exactly divides the given vectors into two classes. Below is a simplified version of the theorem behind SVMs. Given a set of training data

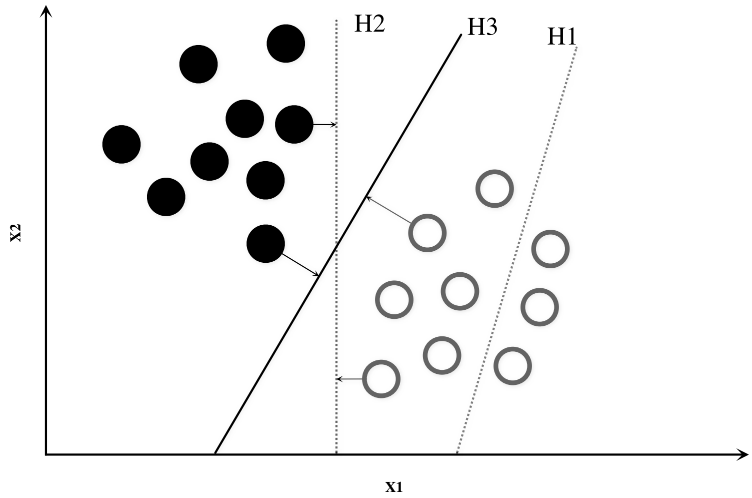

FIGURE 3.5: An example of how SVM chooses hyperplanes based on the optimisation problem (maximum margins) presented below. H1 fails to separate the classes, H2 separates the classes with a small margin and H3 separates the classes with the largest (maximum) margin. This figure has been adapted from that seen in [61]

$\mathcal{D}$ of data points x_1 to x_n, where each class membership c_i is either -1 or 1:

$$\mathcal{D} = \{(x_i, c_i) | x_i \in \mathbb{R}, c_i \in \{-1, 1\}\}_{i=1}^{n}$$

A Support Vector Machine aims to find the maximum margin hyperplane that divides the points $y_i = 1$ from $c_i = -1$. A hyperplane is defined as the set of points x satisfying $w \cdot x - b = 0$, where b, the number of hyperplanes (in this case $b = 2$), and w, a normal vector, are chosen such that the following two parallel hyperplanes are as far away as possible from the middle (dividing hyperplane):

$$w \cdot x - b = -1 \text{ and } w \cdot x - b = 1$$

To prevent any data from falling into the margin, the constraint $y_i(w \cdot x - b) \geq 1$ is added, which leads to the following optimisation problem:

$$\text{minimise } ||w|| \text{ in } (w, b) \text{ subject to } c_i(w \cdot x - b) \geq 1 \text{ (for any } i = 1, ..., n)$$

Since it is difficult to solve $||w||$, the norm of w, as it involves a square root, by mathematical convenience $||w||$ is substituted with $\frac{1}{2}||w||^2$ without changing the solution. Without going into further details, the result is a constraint problem is represented below. The Lagrangian multiplier, α_i, is used as a multiplier of vector x_i, and thus becomes a quadratic programming problem, solved using standard techniques:

$$\min_{w,b} \max_{\alpha} \{\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [c_i(w \cdot x_i - b) - 1]\}$$

However, this solution is only suitable for data that is linearly separable. As is the case in many data mining tasks, and particularly in natural language, vectors are rarely easy to cluster linearly (such as in Figure 3.6). Thus, SVM must handle the cases where the data is not linearly separable, there are more than two predictor variables, and classifications with more than two categories. Much of this work is done through an elegant mathematical trick known as the kernel function, which allows the SVM to map the data into a different dimensional space where a new hyperplane can be used to separate the data. Our original points lie within a two-dimensional space, within an x-y plane. The kernel function takes these original (x, c) coordinates and projects them (according to certain rules) into a much higher (even infinitesimal) dimensional space, demonstrated in Figure 3.7. Indeed, as was the case in Figure 3.6, some

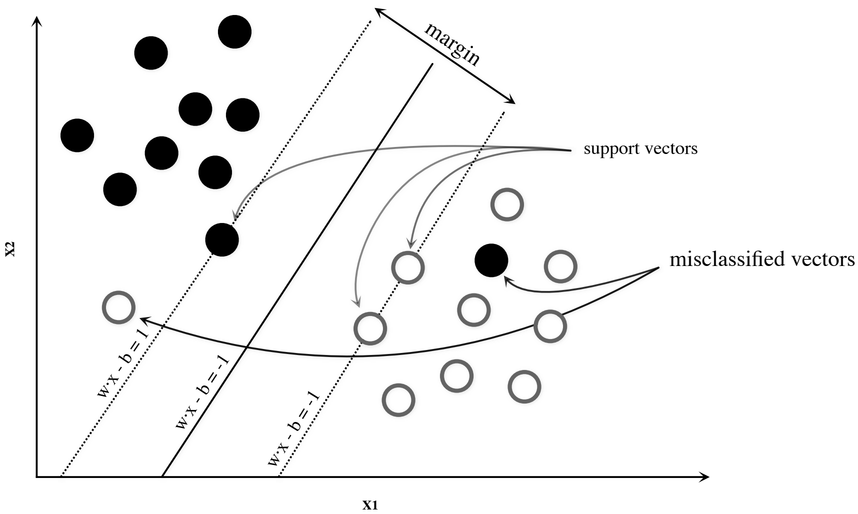

FIGURE 3.6: An example of the chosen hyperplane based on the optimisation problem on a data set. As you can see, the misclassified vectors show it is impossible to linearly separate this data. This figure has been adapted from that seen in [61]

data may not be separable and subject to penalty error or slack cost (given by the trade-off parameter *C*). For a more extensive theorem of Support Vector Machines, see [60] or [61]

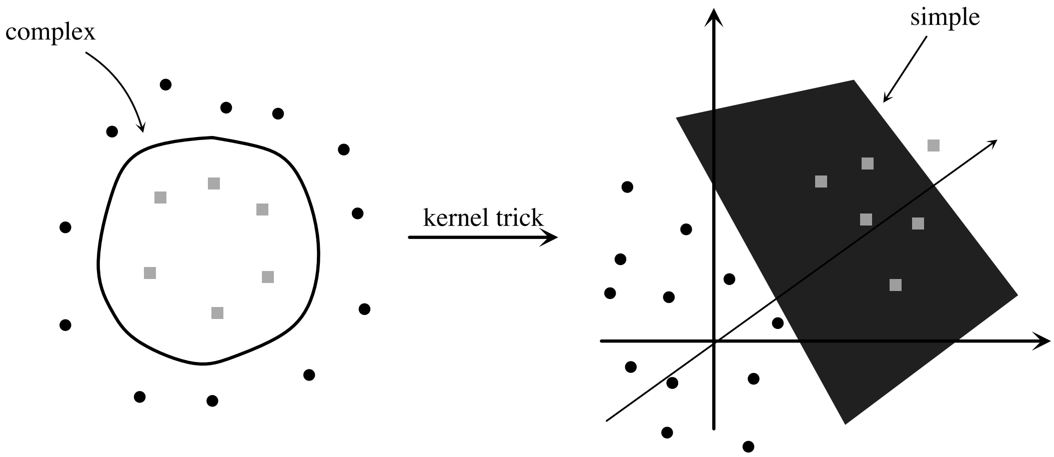

FIGURE 3.7: An example of reducing a complex division in a low dimensions via the kernel function, mapping it to a high dimensionality with a simple (linear) dividing hyperplane. This figure has been adapted from that seen in [61]

3.4.3.2 Multiclass Support Vector Machines

Pure SVMs are only directly applicable to two class (binary) problems, and thus numerous attempts have been made to expand this to a multiclass task through a variety of optimization problems. The dominating approach for Multiclass SVMs is to extend a multiclass problem into many binary classification problems, through one-versus-all or one-versus-one strategies. The training data is split based on the strategy and reduced to a binary problem, with results combined in a voting strategy such as *'winner-takes-all'* or *'max-wins'*. In this treatise, $\text{SVM}_{\text{multiclass}}$ is used, which is a flavour of structural SVM. Unlike other popular approaches, which typically decompose a multiclass problem into multiple independent binary classification tasks, $\text{SVM}_{\text{multiclass}}$'s notion of margin yields a direct method for training multiclass predictors, and achieving state-of-art accuracy in a wide variety of tasks [62].

3.4.4 Classifier Combination

Classifier Combination has proved an effective tool to improve a system's overall performance in many pattern recognition applications, including document classification [63]. In this treatise, a simple 'majority votes' combination strategy is employed, whereby if two classifiers have classified an article as a particular class by majority, that class is chosen. If there is no agreement between all three classifiers, the classifier with the highest F-Score's annotation is used.

3.5 Overall Process

The overall experimental process involves 3 main steps, Pre-Processing, Sentiment Analysis and Machine Classification, as shown in Figure 3.8. Firstly, the company news article is tokenized at the sentence level, then passed into the C&C Parser. The C&C Parser outputs XML data with tokenised words, POS Tags and Named Entities, which is used as input into the sentiment analysis engine. Depending on the experimental feature set, features are identified and included in the final feature set, including the use of Gazetteers, WordNet senses, SentiWordNet, Feature selection such as Entropy, and contextual windows. This output is then prepared for each machine classifier, which requires several different data formats to train the model and evaluate upon. From this final step, the predictions to the original article are given and evaluated against a gold standard - producing result tables given in Chapter 4.

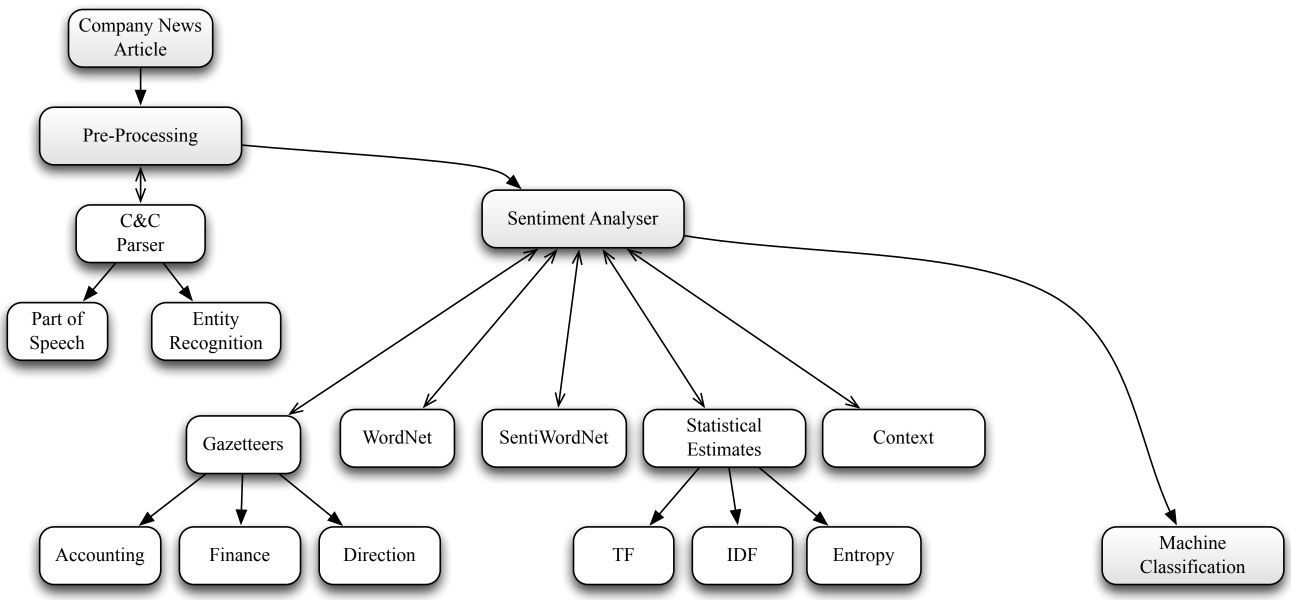

FIGURE 3.8: The Overall Sentiment Analysis Process

3.5.1 System Optimisations

Several optimisations were made to the system in order to improve the speed of the experimental process. As the AFR corpus is a relatively large data set, and some of the pre-processing steps are computationally expensive, optimisations were necessary in order to achieve practical speed for the systems industrial applications aswell as during the design process. The C&C Parser was ran as a SOAP (Simple Object Access Protocol) service, which runs on-line and processes articles through XML-based communication. This sped up the pre-processing step considerably, in particular POS Tagging and Named Entity Recognition, allowing the C&C Parser to load its trained model once only, instead of dumping and reloading for each article. As SentiWordNet is packaged as a flat-file database, optimisations were necessary to make look-ups more affordable (as each word in each article had to retrieve a SentiWordNet score). To speed up this process, the database was loaded into a hash-table, with word-id as the key-value pair, effectively making this search linear ($O(n)$) time. In addition, parsing the XML output from the C&C Parser at run-time proved particularly expensive over large sets of data, so for experimental purposes each article object was serialised after the feature extraction stage. This process is known as 'pickling' in the Python Language, and serialises an object into a flat file which can be loaded directly into memory, aiding experimentation on the development set. Following these optimisations, a ten-fold increase in speed was achieved over the original system design, which now averages 100 articles per second.

3.6 Trading Simulation

A simplistic trading simulation was run in order to view the effects of news sentiment on real stock data. A concise algorithm was written to enable the simulator to trade off such information, by performing both long (buy) and short (sell) trades on ASX equities over the period 1995-2006.

3.6.1 Assumptions & Data

In order to design a trading simulator, several assumptions were necessary in order to make effective judgements on trading direction and visible price effects, the first being that of unlimited funds. This was necessary as for each news event, a trade must be opened with an initial investment, kept constant at $10,000 AUD. In order for this simulation to represent ‘real world’ scenarios, trading costs are also included, at 0.3% of the final value of the investment. This represents a ’round-trip’ of the bid-ask spread, and has been taken from empirical estimations for institutional trades in [64]. In addition, the assumption that any stock can be short-sold is made, however typically they are only available for larger capitalisation shares, provided there is no ban on short selling in the market. No quantitative risk models were employed, as the desired results are based purely on the news event and no other information. The trading position taken by the simulator is based upon the polarity of the news article, where a positive article produces a long position, a negative article produces a short position, and any news event of opposing polarity causes a trade to be closed.

Stock price data was kindly provided by SIRCA, Australian Equities Research. Daily stock price information was sourced for the period 1995-2006 for each of the 1046 companies within the corpus. For each stock quote, the Volume Weighted Average Price (VWAP) was taken as the price the trade could be executed at on that particular day. Informally, VWAP represents the average price paid per share, weighted by the volume traded. Achieving VWAP (or close to) is a reasonable assumption for most automated trading scenarios, as shown in [65]. This is also realistic as each AFR news article is typically published in the morning before market open, and hence using the sentiment prediction trading can be executed at the days VWAP to enter or exit a trade. A corpus of this size lead to several implementation issues, namely incomplete data. Several stocks over this time period went into trading halts, leading to unavailable quotes even if news events occurred that day. Trivially, if no price information was available on the day, the stock cannot be traded upon although the system had provided a prediction. Similarly, if a news article occurred on a public holiday or weekend, the prediction could only be enacted on the following trading day (to account for this, the simulator waits up to 3 days to book a trade).

3.6.2 Algorithm

This short algorithm was written for the trading simulator, shown in Figure 3.9, which takes the predictions made by the Sentiment Analysis System, and trades upon this information. Each article is processed and classified, with the trading simulator using the combined ten-fold predictions as input. The algorithm is also run using the corpus' annotations, to compare performance of the annotator's classification and what the system believes it to be. The algorithm runs for each trading day, and within that each article for that day, and acts upon this information. Depending on whether a long or long/short strategy is used, it opens a trade based on the classified sentiment of an article, and will close trades based on three situations - an article of opposing polarity is encountered, the return is above 15%, or the trade expires based on the given trading window. A trade may be kept open for another period if additional news is found of the same polarity, or if its return is currently above 5%.

```
trading_window = 7 days
for each day in trading period:
    for each open trade:
        if trading_return > 15%:
            close trade
        elif trading_return > 5%:
            extend trading period by trading_window
    for each article in the day:
        article_polarity = { pos, neg, neu }
        if article_polarity != neu:
            if there is no open trade in company:
                open a new trade(article_polarity)
            else there is currently an open trade in company:
                if trade_polarity != article_polarity:
                    close trade
                else:
                    extend trading period by trading_window
```

FIGURE 3.9: This short trading algorithm, written in pseudo-code, represents what the trading simulator runs for each day of the investment period, primarily acting upon news event outcomes predicted by the Sentiment Analysis System

CHAPTER 4

Analysis

4.1 Evaluation Measures

4.1.1 Precision & Recall

Evaluation in Information Retrieval makes frequent use of the notions of precision and recall, and make for an easy extension into the language-processing domain. An arguably complete measure of a system's performance is given by the precision-recall curve, and the balanced F-Score [66]. True Positives (tp_c) are defined as the number of correctly assigned articles to a particular sentiment class (c), while True Negatives (tn_c)) account for the number correctly rejected documents. False Positives (fp_c) are the number of incorrectly assigned documents to class (c), and False Negatives (fn_c) are the number of incorrectly rejected documents. Statistically, a False Positive is a Type I error (by rejecting the null hypothesis given that it is actually true), and a False Negative is a Type II Error (by failing to reject the null hypothesis given that the alternative hypothesis is actually true). Precision (P_c) is defined as the probability that a document is relevant, given that it is returned by the system (ie. the proportion of selected documents the system got right), while Recall (R_c) is the probability that a relevant document is actually returned (ie. the proportion of the target documents that the system selected). They are given by the formulae:

$$
\begin{aligned}
P_c &= \tfrac{tp_c}{tp_c+fp_c} \\
R_c &= \tfrac{tp_c}{tp_c+fn_c}
\end{aligned}
$$

4.1.2 Traditional & Modified F-Score

The F-Score is common in language processing and other information retrieval experiments due to the convenience of which it can wholly describe a systems' overall performance in a single figure [67].

$$FScore_c = \frac{2 \times P_c \times R_c}{P_c + R_c}$$

In this paper, results are reported with both a traditional harmonic mean F-Score and a Modified F-Score. A novel modified F-Score has been developed for this type of evaluation, as for the purposes of a trading algorithm, misclassified articles have varied opportunity cost. When an article is annotated as positive (or negative), and is classified as neutral, a false positive is counted. The argument here is that this misclassification represents a 'missed opportunity' for the trading scenario, and therefore should not be penalised in the evaluation. The false positives (of the form $annotation \rightarrow prediction$), $positive \rightarrow neutral$ and $negative \rightarrow neutral$ are discarded as a result of this novel approach. Given the difficulty of sentiment analysis, it is believed that this is a fair adjustment in evaluation.

4.1.3 Confusion Matrices

In predictive analysis, Confusion Matrices are useful to help visualise the number of examples from one class classified into another (or the same) class. It is an extension of a contingency table, to help show how many classes are being misclassified based on the gold standard data.

		Actual Value	
		p	**n**
Predicted	**p**	True Positive	False Positive
Value	**n**	False Negative	True Negative

TABLE 4.1: An example of a Confusion Matrix with two classes, positive & negative

Confusion matrices are especially important when a data set is 'unbalanced', such as the AFR corpus, with the majority of annotations being positive. Confusion matrices in this treatises are set out in the form show in Table 4.1.

4.1.4 Financial Return

Financial Return is used to evaluate the trading simulator's performance, by using the traditional definitions of Profit/Loss and Return. P_{close} and P_{open} represent opening and closing prices, which by assumption, are the Volume Weighted Average Prices (VWAP) of that day. The trading cost variable (C) is cost associated with each transaction, which include a round-trip of the bid-ask spread (fixed at 0.3%). As share units (U) are unaltered during trading, it is moved to the outside of the equation:

$$Profit/Loss = (1 - C) \times (P_{close} - P_{open}) \times U$$

$$Return = (1 - C) \times \frac{P_{close} - P_{open}}{P_{open}}$$

4.2 Development & Test Corpora

As is typical in many NLP tasks, the corpus was sectioned into a development and a test set, in which the experimental process focuses solely on achieving the best possible results for a development set, which are then applied to a 'held-out' test set. This removal of part of the corpus represents a 'real-world' evaluation, to see how well a system performs on completely unseen data. In addition, due to the cost of annotation, the corpus is sectioned into Single Annotated and Double Annotated sets.

4.2.1 Ten-fold Cross Validation

Cross Validation or Deleted Estimation is a process applied in the development stage in which each part of the development data is used both as initial training data and as 'held-out' test data. Holding out part of the data as a test set helps determine how well the feature set generalises over a corpus. In Ten-fold Cross Validation, the data is split into ten subsets (10% vs. 90%) of approximately equal size. The system is then trained ten times, using the omitted subset of data as evaluation, helping estimate how accurately a predictive model will perform in practice, and which are the most informative features. Ten-fold Cross validation is the preferred method for many information retrieval experiments, both in terms of a lower variance and smaller bias, in comparison to similar methods such as boot-strapping and held-out estimation [68]. Since 10 predictive models are relatively expensive to train for 3 separate classifiers, training time is reduced by restricting the number of iterations each classifier performs generating the model.

4.3 Results

This section presents the system results based on the evaluation measures discussed in Section 4.1, namely Precision, Recall, F-Score and Trading Return. Results are sectioned based on the Development/Test split, and Trading Simulation results, and are abbreviated versions of the tables found in Appendices A & B.

4.3.1 Development

4.3.1.1 Single Annotations

The Single Annotated Development results for the three classifiers - $\text{SVM}_{\text{multiclass}}$, MegaM and Naïve Bayes are displayed in Table 4.2. The baseline for these results is a Unigrams feature set, in which one chooses the most frequently occurring words in the corpus (described in Section in 3.2.4), and is a typical baseline for document classification experiments. The baselines for $\text{SVM}_{\text{multiclass}}$, MegaM and Naïve Bayes are 49.53%, 53.71% and 53.95% respectively. Each feature extraction/selection technique is progressivly applied to the development set, with each F-Score reported. Emphasized in bold are the best results for each classifier in both Traditional and Modified F-Scores. For the sake of brevity, the size of each feature set is only included in the following Table 4.3. Table 4.2 shows that, with each successive technique applied, F-Scores are seen to generally improve. From these figures, it can be seen the with each classifier, the baseline is beaten by 8.5%, 4.46% and 7.83% for $\text{SVM}_{\text{multiclass}}$, MegaM and Naïve Bayes, signally that these linguistically motivated techniques are moving in the right direction. The biggest improvements against the baseline are generally seen with the introduction of the contextual window and part of speech tags. Interestingly, the modified F-Score results

	SVM$_{multiclass}$		MegaM		Naïve Bayes	
Feature Set	Traditional	Modified	Traditional	Modified	Traditional	Modified
Unigrams	51.47%	63.63%	56.48%	64.54%	56.78%	64.12%
+Ngrams	51.62%	64.23%	55.46%	64.33%	56.66%	63.56%
+Accounting Gazett.	52.32%	67.44%	54.33%	65.67%	54.94%	67.31%
+Finance Gazett.	54.34%	71.51%	56.33%	68.60%	59.56%	72.74%
+Window=1	56.22%	71.51%	**58.17%**	**69.52%**	60.51%	**73.57%**
+POS Tag	**58.03%**	72.42%	57.57%	65.93%	**61.78%**	72.50%
+SWN, Direction Gazett.	57.62%	72.49%	57.62%	66.95%	61.47%	72.13%
+Entropy	57.62%	72.49%	57.26%	66.74%	61.47%	72.13%
+IDF	57.63%	**72.50%**	57.49%	66.62%	61.47%	72.13%

TABLE 4.2: Traditional and Modified F-Score results on the Development Set (Single Annotated), with 5424 articles, averaged over ten-folds

for SVM$_{multiclass}$ continually increase with the addition of more features. This suggests, from the definition of the modified F-Score, that this is leading to an improvement in classification of positive and negatively annotated articles. The feature selection techniques added at these stages increase the number of features used to develop the model, as shown in the next table, Table 4.3, suggesting that the system is producing more but slightly different features.

The Single Annotated Development results for classifier combination can be found in Table 4.3. Using Combined Classifiers, one saw a continual improvement in F-Score against the baseline by adding each feature set, leading to the baseline being beaten by 7.84% and 6.31% in Traditional & Modified F-Score respectively. However, like the other 3 classifier results, the highest improvements came from the addition of the Financial Gazetteer, Contextual Window and Part of Speech tags. For the sake of brevity, only the F-Scores are displayed in this chapter, and the full Precision/Recall/F-Score tables can be found in the Appendix A.

		Combined	
Feature Set	**Features (No.)**	Traditional	Modified
Unigrams	474	53.77%	66.83%
+Ngrams	3314	55.51%	67.58%
+Accounting Gazett.	3338	55.39%	67.65%
+Finance Gazett.	12979	59.35%	72.94%
+Window=1	13105	60.48%	**73.14%**
+POS Tag	13105	61.60%	72.13%
+SWN, Direction Gazett.	13201	61.53%	72.31%
+Entropy	13199	61.54%	72.34%
+IDF	13184	**61.61%**	72.29%

TABLE 4.3: Traditional and Modified F-Score results on the Development Set (Single Annotated) by combining classifiers, with 5424 articles, averaged over ten-folds

4.3.1.2 Double Annotations

The Double Annotated Development results for the three classifiers - $\text{SVM}_{\text{multiclass}}$, MegaM and Naïve Bayes are displayed in Table 4.4. These results are based on the smaller subset of data, of double annotations, by only using articles that both annotators agreed on. The baseline for these results is the Unigram feature set, with each successive feature set applied, and both Traditional and Modified F-Scores reported. The size of each feature set is displayed in the following Table 4.5. With each progressive feature set applied, one can see that the baseline is beaten by 6.75%, 3.93% and 3.24% for $\text{SVM}_{\text{multiclass}}$, MegaM and Naïve Bayes respectively. These results tend to be not as strong as the improvements shown in the Single Annotated experiments in Table 4.2, which is unexpected. However, due to a smaller sample size and training/evaluation split, it is believed that the machine learners suffered from a lack of training data. The biggest improvements were seen with the introduction of the contextual windows and POS tags, similar to the Single Annotated results, with MegaM achieving the best results.

	$\text{SVM}_{\text{multiclass}}$		MegaM		Naïve Bayes	
Feature Set	Traditional	Modified	Traditional	Modified	Traditional	Modified
Unigrams	51.47%	63.63%	56.48%	64.54%	56.78%	64.12%
+Ngrams	51.62%	64.23%	55.46%	64.33%	56.66%	63.56%
+Accounting Gazett.	52.26%	64.87%	55.75%	64.91%	56.43%	63.50%
+Finance Gazett.	57.70%	68.17%	59.64%	69.44%	**60.02%**	**69.44%**
+Window=1	58.20%	68.64%	60.40%	70.04%	60.01%	69.13%
+POS Tag	**58.22%**	**68.70%**	**60.41%**	**70.04%**	60.01%	69.13%
+SWN, Direction Gazett.	56.89%	68.00%	57.41%	68.24%	59.59%	68.95%
+Entropy	56.96%	68.06%	57.46%	68.77%	59.96%	69.02%
+IDF	56.99%	68.11%	57.43%	68.53%	59.73%	68.94%

TABLE 4.4: Traditional and Modified F-Score results on the Development Set (Double Annotated), with 1177 articles, averaged over ten-folds

		Combined	
Feature Set	**Features (No.)**	Traditional	Modified
Unigrams	471	56.13%	67.42%
+Ngrams	2424	56.83%	64.97%
+Accounting Gazett.	2756	56.90%	64.93%
+Finance Gazett.	8261	58.90%	69.06%
+Window=1	8438	60.22%	69.90%
+POS Tag	8441	**60.22%**	**69.91%**
+SWN, Direction Gazett.	8512	58.65%	68.92%
+Entropy	8438	58.90%	69.39%
+IDF	8417	58.86%	69.28%

TABLE 4.5: Traditional and Modified F-Score results on the Development Set (Double Annotated) by combining classifiers, with 1177 articles, averaged over ten-folds

The Double Annotated Development results for classifier combination are displayed in Table 4.5, along with the sizes of each feature set. The baselines here of 56.13% and 67.42% were beaten by 4.09% and 2.49% for Traditional and Modified F-Score respectively. This saw less improvement in using classifier combination than the previous Single Annotated experiments, and can be explained by the small size of the development data.

4.3.2 Test

Part of the corpus has been held out from the development stage, in order to represent a 'real-world' evaluation, to see how well a system performs on completely unseen data. Here the machine learners learn a model based on the whole development set, and then are evaluated on this held out data.

4.3.2.1 Single Annotations

The Single Annotated Test Results are presented for all 4 classification methods in Table 4.6, with both Traditional and Modified Precision, Recall & F-Score. Each classifier trains its model from the development set of 5424 articles, and is evaluated on the held out set of 863 articles. The feature set which gave the best performance in the development step is used, that is - Ngrams, Accounting & Finance Gazetteers, Contextual Window (1) and Part of Speech Tags. All 4 classifiers improve on their performance in the development data set which is typical for two reasons, as they are not averaged over ten-folds and that the machine learners benefit from larger training data. Classifier combination is the highest performer, beating $\text{SVM}_{\text{multiclass}}$ and MegaM by up to 8% in Traditional F-Score. However, there is only a 1-2% improvement in the Modified F-Score, which shows that what is gained by Classifier Combination is an improvement in the distinction between positive or negative articles with neutral ones, as seen in the Confusion Matrix in 4.7 with significant amounts of $positive \rightarrow neutral$ and $negative \rightarrow neutral$ misclassifications. The large improvement in Traditional F-Score is primarily due to the effectiveness of Naïve Bayes in making this distinction, and shows it robustness in the face of more complex learning routines.

	Traditional			Modified		
Method	Precision	Recall	F-Score	Precision	Recall	F-Score
$SVM_{multiclass}$	60.07%	58.98%	59.52%	75.07%	73.49%	74.27
MegaM	61.32%	63.27%	62.28%	73.43%	69.52%	71.42
Naïve Bayes	**66.82%**	67.32%	67.07%	**77.67%**	73.12%	75.33
Combined	66.36%	**68.02%**	**67.18%**	77.59%	**74.21%**	**75.86**

TABLE 4.6: Traditional & Modified Precision, Recall and F-Score for all 4 classification methods, trained on 5424 articles and tested on the held out data set of 863 articles, with 42999 features

		Actual Value		
		Positive	**Negative**	**Neutral**
Predicted Value	**Positive**	386	40	91
	Negative	40	134	42
	Neutral	36	27	67

TABLE 4.7: The Confusion Matrix for the single annotated test data set (863 articles), using the Combined Classifiers. This shows the system results for our predicted values against the annotations.

4.3.2.2 Double Annotations

The Double Annotated Test Results are presented for all 4 classification methods in Table 4.8, with both Traditional and Modified Precision, Recall & F-Score. Each classifier has trained a model based on the development set of 1177 articles (producing 9275 features), and is then evaluated on the held out data set of 118 articles. As with the single annotated test results, Ngrams, Accounting & Finance Gazetteers, Contextual Window (1) and Part of Speech Tags are used as the test feature set. In this instance, a huge improvement in performance is seen, from 60.22% in the development set up to 76.54% for Traditional F-Score, and is attributable to the lack of adequate training data when using ten-fold cross validation. The large improvements

	Traditional			Modified		
Method	Precision	Recall	F-Score	Precision	Recall	F-Score
$SVM_{multiclass}$	75.84%	73.73%	74.77%	84.64%	84.31%	84.48%
MegaM	74.21%	68.64%	71.32%	83.92%	84.14%	84.03%
Naïve Bayes	77.03%	75.42%	76.22%	84.66%	82.54%	83.59%
Combined	**77.69%**	**75.42%**	**76.54%**	**86.23%**	**85.77%**	**86.00%**

TABLE 4.8: Traditional & Modified Precision, Recall and F-Score for all 4 classification methods, trained on 1177 Double Annotated articles and tested on the held out data set of 118 articles, with 9275 features

		Actual Value		
		Positive	**Negative**	**Neutral**
Predicted Value	**Positive**	64	5	2
	Negative	5	14	4
	Neutral	8	4	12

TABLE 4.9: The Confusion Matrix for the double annotated test data set (118 articles), using the Combined Classifiers. This shows the system results for our predicted values against the annotations.

seen from development into test could also be explained by the high percentage of positive articles found in the test data, as seen in Table 4.9. The system was found to be typically very strong in classifying positive articles, due to the percentage of positive articles (45%) found in the corpus overall. When comparing the two tables, Single Annotated in Table 4.6 and Double Annotated in Table 4.8, one can see that there has been a significant reduction in the variability of each classifier. Indeed this can be seen as a result of the agreement between annotators creating a more robust training model for the more complicated machine learning algorithms $SVM_{multiclass}$ and MegaM, with Naïve Bayes only outperforming to two by 2-4% in Traditional F-Score.

4.3.3 Trading Simulation

The trading simulation was run for all predictions made during the Single Annotated Development experiments by using the output of the Combined Classifiers. Each simulation takes approximately 2 minutes to run and output its trading results (due to the loading of over 2 million rows of stock tick data). The simulation results for each year are displayed graphically in Figure 4.1, with two different trading windows and a long strategy, 7 and 90 days. For each year, each trading return is averaged by the number of trades made in that year. Using the 7 day trading window, primarily positive returns were achieved, with an annual average return of 0.76%, with the year 2000 having the only negative return. Similarly using a 90 day trading window produces a large dip in returns in the year 2000, however averaging a significantly higher annual return of 4.27%. Certainly, this dip in trading performance in 2000 can be attributed to volatility in the overall equity market in that year. Interestingly, the correlation seen between the number of trades made and positive returns was 29.04% for the 7 day window and 26.80% for the 90 day window, following the general trend of the All Ordinaries index over the time period 1995 to 2006, seen previously in Figure 3.1. The more active the trading simulation was, the more positive returns, showing its potential with larger amounts of data. However, in terms of the magnitude of positive returns, a less frequent trading strategy (such as the 90 day window) produced significantly higher returns. Indeed, if one trades less frequently, they will avoid high trading costs, and profit more from the swing or momentum of a stock following an information release. Table 4.10 shows the trading return for the simulator based on each individual industry and similarly averaged over the number of trades for both 7 day and 90 day trading windows. The system performed consistently well in several industries (in both

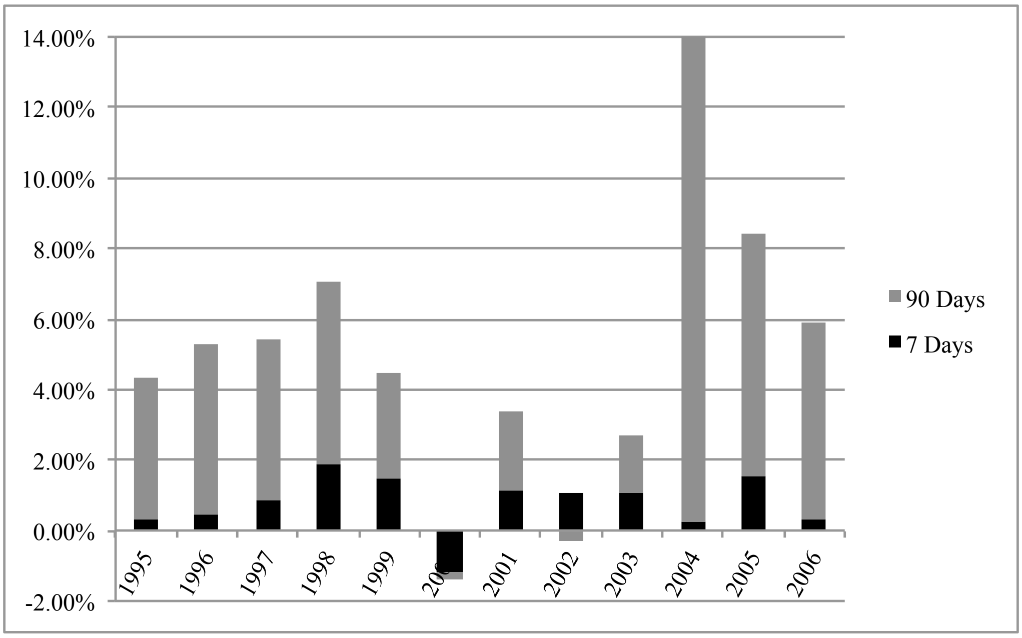

FIGURE 4.1: Yearly Trading Results using a 7 and 90 day trading window, with no short-selling.

trading windows), including Banks, Consumer Services, Diversified Financials, Materials and Real Estate. Areas in which it did not perform well are Capital Goods and Media in particular, which both saw negative returns for each strategy. Interestingly, a change from the 7 day to 90 day trading window saw an increase of over 28% return in the Transportation industry. These figures also confirm that a less frequently traded strategy result in a much higher, more consistent positive return. Further tables are available in Appendix B, with a full trading summary for the 7 day trading window, as well as long/short strategies for both 7 and 90 day windows. Using a long/short strategy, as shown in Tables B.4 and B.5, degrades the trading performance slightly, averaging yearly returns of 0.57% and 3.15% for 7 day and 90 days respectively. One reason for this is that negative news does not have as deep an impact on price when the general trend of the market index tends to increase, and this is where Beta hedging using index futures may have proved useful to isolate the impact of the news on the stock price.

	7 days		90 days	
Industry	Trades	Return	Trades	Return
#N/A	33	-0.68%	25	0.24%
Automobiles & Components	3	4.55%	2	2.24%
Banks	143	0.46%	80	5.30%
Capital Goods	9	-4.78%	9	-2.94%
Commercial & Professional Services	1	0.44%	1	17.68%
Consumer Durables & Apparel	9	-0.56%	7	6.73%
Consumer Services	36	1.68%	28	6.02%
Diversified Financials	57	1.29%	36	7.49%
Energy	7	0.87%	7	3.99%
Food & Staples Retailing	13	0.27%	11	1.42%
Food, Beverage & Tobacco	10	1.79%	9	4.44%
Health Care Equipment & Services	6	2.34%	6	5.98%
Insurance	73	0.29%	41	1.18%
Materials	178	0.65%	129	2.29%
Media	27	-1.98%	23	-0.43%
Pharmaceuticals, Biotechnology & Life Sciences	5	5.63%	5	4.72%
Real Estate	43	4.65%	30	6.92%
Retailing	12	-1.20%	9	0.96%
Software & Services	4	1.52%	4	15.34%
Telecommunication Services	29	2.69%	23	0.47%
Transportation	20	-0.33%	16	27.87%
Utilities	37	0.45%	19	6.44%

TABLE 4.10: Trading Results based on Industry for both 7 day and 90 day windows

4.4 Discussion

Both the development and test experimental results produce some very encouraging F-Scores, by beating the baseline by almost 8%. Adding the Financial Gazetteer, Part of Speech Tags and Contextual windows saw the biggest improvement against the baseline of the Unigram feature set (averaging 1.5-2.0% for each classification method against the previous feature set).

Adding the SentiWordNet feature set found a slight degradation in Modified F-Score, however this decrease was not as sharp as in Traditional F-Score. Though initially disappointing, considering its wide usage in Sentiment Analysis, further investigation into the SentiWordNet corpus returned some interesting results. Within the corpus, only a small percentage of words have actual polarity greater than zero, and the majority of words that had positive polarity scores were adjectives and adverbs, which were accounted for in a previous feature set. As a feature selection technique, SentiWordNet does not have the same descriptive power as simply using Part of Speech tags and including adjective/adverbs in the machine learning step. Development results for the double annotated articles didn't see as much improvement as the Single Annotated articles, and this is believed to be because of a lack of training data. Since the agreement score (Kappa) was 70.64% for the smaller subset of articles, the development scores seemed to be hurt by a lack of training data (from 5424 articles down to 1177 articles) and the ten-fold cross-validation split. The small set of double annotated was a result of the costly annotation process, however the benefits seen from double annotated data can be seen once the test experiments were run. This saw a huge boost in Precision, Recall & F-Scores, which is a very promising result using the larger set of training data. The Trading Simulation results were performed using 7 day and 90 day trading windows, using the novel algorithm with both long and long/short strategies. The positive annual trading returns of 0.76% (7 day) and 4.27% (90 day) both confirm the weak-form Efficient Market Hypothesis (EMH), and the practicality of using news sentiment in algorithmic trading scenarios. Financial academics are in continual debate over the existence of EMH, and its impact on stock prices. However, the positive results presented in 4.3.3 support the presence of EMH theory for the period 1995-2006. In particular, when both positive and negative articles were used in a long/short strategy, only a small degradation in return was seen. Shorting a financial instrument is generally seen as a riskier venture [14], so the fact that positive returns are still achieved shows that there is both positive and negative impacts on price following an information release. The large improvement from 7 day to the longer term 90 day trading window shows that there is potential for this to act as a portfolio selection technique, to benefit from longer term swings in price. Trading less frequently not only avoids large trading costs, it also allows for larger profits while reducing losses through the stop-loss order encapsulated in the algorithm (in which a trade a is closed when return dips below a certain percentage). The most profound aspect to this trading simulation is the fact that no other background information is utilised, such as quantitative risk models, financial statement research or any other modelling, to produce positive returns in the equity market.

CHAPTER 5

Conclusion

The tremendous growth in the availability of information in all forms of news, media, blogs and forums has led researchers to investigate and refine methods to cultivate such large amounts of qualitative knowledge. It is not only important to extract this information, but also to be able to exploit and harness such data in practical applications. In this treatise, a vast array of Natural Language Processing techniques have been applied to the domain of Finance in order to capitalise on efficient market theory and to examine how the news influences market behaviour. Experimental procedures have been developed to statistically model financial sentiment (and in essence, investor behaviour), to see the efficacy of news on stock prices in the Australian Equities market (ASX). Although seemingly objective, newspaper articles tend to express an author's sentiment hidden within the language used, through the use of adjectival and adverbial phrases in relation to an information release. The system designed for this treatise uses annotated articles to build a statistical model to recognize and predict the polarity or sentiment orientation of the author and the content within such news. The corpus was specifically constructed and annotated for these experiments, and is in this sense, very unique. Leveraging the knowledge of the financial domain, through the known effects of key indicators such as net income, profit and resources, and linguistically motivated sentiment criteria such as adjectives and adverbs, a system has been constructed to successfully predict stock market price movements. Significant improvements during the development stage of experimentation came in particular from financial dictionaries, contextual features and parts of speech, seeing a baseline of 51.47% F-Score boosted by up to 8%. Four machine learning techniques were applied to see the effects of various statistical models, with Naïve Bayes clearly the strongest. The most promising results were expressed in the test phase of experimenting, using the Double Annotated data set, with F-Scores in the range of 71-76%. These are significant in that they show the potential strength of a predictive model trained on a sufficiently large amount of data annotated by two people in complete agreement. The system presented is also unique in the sense that it utilises these predictions made by the machine classifiers to simulate a trading scenario on real-world stock data, and achieves positive annualised returns of 0.76% (7 days) and 4.27% (90 days). What is possibly even more profound, is that it achieves these returns based on a simplistic algorithm that unambiguously buys shares on positive news, and sells on negative news, without any other inputs or prior knowledge into the decision making process (such as quantitative models or balance sheet research). The results not only support weak-form Efficient Market Hypothesis (EMH) but also show the practicality and potential of a system to algorithmically trade in financial markets, making decisions many orders of magnitude faster than any human could possibly do.

Although some very promising outcomes were produced by the Sentiment Analysis Engine in this treatise, there are certainly ways in which to improve upon these results and further investigate this burgeoning field of research. Since the annotation of a novel corpus proved an extremely costly adventure (in terms of time), much of the corpus was left without annotation, and as such could not be trained or evaluated upon. This would certainly improve the robustness and strength of the predictive models, as well as seeing advancement in accuracy through various feature selection techniques (which suffered from the lack of training data). SIRCA was kind enough to provide the daily stock data for the trading simulation, however historical Beta (β) figures were not available through any academic means. Using historical beta and the Capital Asset Pricing Model (CAPM), the simulator could have effectively separated any general market movements from the change in price, and more thoroughly isolate the effect of the news on the company's share price. This is done through statically (or dynamically) hedging the share by a proportional (i.e. $1/\beta_i$) amount of market index future contracts, thus normalising any returns and finding the true effect of such financial news. Also, the significantly positive returns generated from using a 90 day window show the potential for Sentiment Analysis outside the realm of algorithmic trading, such as creating an online 'Sentiment Index', used to track the general sentiment (i.e. 'Bull' or 'Bear') of individual companies based on collated news. This could be done by assigning polarity scores (-1 to 1) for news articles, and reporting these figures as a single indicator (perhaps as a function of time) for whether a company is currently a buy, hold or sell, or combining this into an overall market sentiment index (and weighting by each companies market capitalisation). An extension to this would also be to have the engine process Reuters global news feeds, tuned with a high Recall accuracy, and have it search for and 'flag' information that may have an impact on a variety of instruments, markets, or even economies. On a global scale, this could have huge potential for hedge funds looking for expanding markets with substantial profits. Through the continual growth and advancement in language processing, statistical learning and machine readable news, there is no doubt that this field of computing will become increasingly crucial to such a competitive financial world.

APPENDIX A

System Results

A1 Single Annotated

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	474	51.84%	47.41%	49.53%	68.73%	63.30%	65.90%
+Ngrams	3314	53.53%	49.93%	51.67%	69.82%	64.64%	67.13%
+Accounting Gazett.	3338	54.07%	50.67%	52.32%	70.14%	64.93%	67.44%
+Finance Gazett.	12979	56.76%	52.12%	54.34%	73.12%	69.97%	71.51%
+Window=1	13105	57.21%	55.27%	56.22%	73.11%	69.97%	71.51%
+POS Tag	13105	**58.66%**	**57.42%**	**58.03%**	**73.85%**	71.05%	72.42%
+SWN, Direction	13201	58.16%	57.09%	57.62%	73.74%	71.27%	72.49%
+Entropy	13199	58.15%	57.10%	57.62%	73.71%	71.20%	72.44%
+IDF	13184	58.18%	57.10%	57.63%	73.76%	**71.28%**	**72.50%**

TABLE A.1: Traditional & Modified Precision, Recall and F-Score for $\text{SVM}_{\text{multiclass}}$, trained and tested using Ten Fold Cross-Validation, with a total of 5424 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	474	54.82%	52.64%	53.71%	69.83%	60.28%	64.62%
+Ngrams	3314	55.46%	53.91%	54.67%	69.83%	61.77%	65.55%
+Accounting Gazett.	3338	55.03%	53.65%	54.33%	69.80%	62.00%	65.67%
+Finance Gazett.	12979	57.05%	55.63%	56.33%	71.97%	65.53%	68.60%
+Window=1	13105	58.42%	**57.92%**	**58.17%**	**72.61%**	**66.68%**	**69.52%**
+POS Tag	13105	**58.56%**	56.62%	57.57%	70.92%	61.61%	65.93%
+SWN, Direction	13201	58.23%	57.02%	57.62%	71.36%	63.05%	66.95%
+Entropy	13199	57.89%	56.65%	57.26%	71.07%	62.90%	66.74%
+IDF	13184	57.97%	57.02%	57.49%	71.04%	62.72%	66.62%

TABLE A.2: Traditional & Modified Precision, Recall and F-Score for MegaM, trained and tested using Ten Fold Cross-Validation, with a total of 5424 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	474	54.89%	53.05%	53.95%	70.29%	64.38%	67.20%
+Ngrams	3314	55.59%	54.35%	54.97%	70.51%	64.15%	67.18%
+Accounting Gazett.	3338	55.58%	54.31%	54.94%	70.58%	64.32%	67.31%
+Finance Gazett.	12979	60.70%	58.46%	59.56%	75.48%	70.20%	72.74%
+Window=1	13105	61.02%	60.01%	60.51%	75.52%	**71.73%**	**73.57%**
+POS Tag	13105	**62.77%**	**60.81%**	**61.78%**	**75.82%**	69.46%	72.50%
+SWN, Direction	13201	62.39%	60.58%	61.47%	75.56%	69.00%	72.13%
+Entropy	13199	62.39%	60.58%	61.47%	75.56%	69.00%	72.13%
+IDF	13184	62.39%	60.58%	61.47%	75.56%	69.00%	72.13%

TABLE A.3: Traditional & Modified Precision, Recall and F-Score for Naïve Bayes, trained and tested using Ten Fold Cross-Validation, with a total of 5424 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	474	54.87%	52.72%	53.77%	70.18%	63.79%	66.83%
+Ngrams	3314	56.26%	54.79%	55.51%	70.88%	64.57%	67.58%
+Accounting Gazett.	3338	56.11%	54.69%	55.39%	70.89%	64.69%	67.65%
+Finance Gazett.	12979	60.22%	58.50%	59.35%	74.99%	70.99%	72.94%
+Window=1	13105	60.51%	60.45%	60.48%	74.92%	**71.45%**	**73.14%**
+POS Tag	13105	**61.90%**	61.30%	61.60%	75.05%	69.43%	72.13%
+SWN, Direction	13201	61.75%	61.32%	61.53%	75.13%	69.70%	72.31%
+Entropy	13199	61.79%	61.29%	61.54%	**75.16%**	69.73%	72.34%
+IDF	13184	61.80%	**61.43%**	**61.61%**	75.13%	69.66%	72.29%

TABLE A.4: Traditional & Modified Precision, Recall and F-Score with Classifier Combination, trained and tested using Ten Fold Cross-Validation, with a total of 5424 articles

A2 Double Annotated

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	471	52.75%	50.25%	51.47%	66.65%	60.86%	63.63%
+Ngrams	2424	52.59%	50.68%	51.62%	67.35%	61.38%	64.23%
+Accounting Gazett.	2756	53.66%	50.93%	52.26%	68.26%	61.81%	64.87%
+Finance Gazett.	8261	58.83%	56.61%	57.70%	70.91%	65.63%	68.17%
+Window=1	8438	59.15%	57.29%	58.20%	71.47%	66.03%	68.64%
+POS Tag	8441	**59.18%**	**57.29%**	**58.22%**	**71.49%**	**66.12%**	**68.70%**
+SWN, Direction	8512	57.52%	56.27%	56.89%	70.51%	65.66%	68.00%
+Entropy	8438	57.61%	56.32%	56.96%	70.58%	65.71%	68.06%
+IDF	8417	57.65%	56.34%	56.99%	70.61%	65.79%	68.11%

TABLE A.5: Traditional & Modified Precision, Recall and F-Score for $SVM_{multiclass}$, trained and tested using Ten Fold Cross-Validation, with a total of 1177 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	471	56.70%	56.27%	56.48%	68.25%	61.21%	64.54%
+Ngrams	2424	55.09%	55.85%	55.46%	67.21%	61.70%	64.33%
+Accounting Gazett.	2756	55.40%	56.10%	55.75%	67.78%	62.27%	64.91%
+Finance Gazett.	8261	61.22%	58.14%	59.64%	72.56%	66.58%	69.44%
+Window=1	8438	**61.80%**	59.07%	60.40%	73.29%	67.06%	70.04%
+POS Tag	8441	61.79%	**59.10%**	**60.41%**	**73.29%**	**67.06%**	**70.04%**
+SWN, Direction	8512	58.79%	56.09%	57.41%	71.43%	65.32%	68.24%
+Entropy	8438	58.92%	56.08%	57.46%	71.66%	66.10%	68.77%
+IDF	8417	58.80%	56.12%	57.43%	71.02%	66.20%	68.53%

TABLE A.6: Traditional & Modified Precision, Recall and F-Score for MegaM, trained and tested using Ten Fold Cross-Validation, with a total of 1177 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	471	56.52%	57.03%	56.78%	67.68%	60.91%	64.12%
+Ngrams	2424	56.80%	56.53%	56.66%	66.34%	60.99%	63.56%
+Accounting Gazett.	2756	56.59%	56.27%	56.43%	66.27%	60.96%	63.50%
+Finance Gazett.	8261	61.09%	**58.98%**	**60.02%**	72.52%	**66.62%**	**69.44%**
+Window=1	8438	61.81%	58.31%	60.01%	**72.94%**	65.69%	69.13%
+POS Tag	8441	**61.81%**	58.31%	60.01%	72.94%	65.69%	69.13%
+SWN, Direction	8512	61.22%	58.05%	59.59%	72.80%	65.49%	68.95%
+Entropy	8438	61.37%	58.61%	59.96%	72.84%	65.58%	69.02%
+IDF	8417	61.20%	58.32%	59.73%	72.79%	65.48%	68.94%

TABLE A.7: Traditional & Modified Precision, Recall and F-Score for Naïve Bayes, trained and tested using Ten Fold Cross-Validation, with a total of 1177 articles

		Traditional			Modified		
Feature Set	**Features (No.)**	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigrams	471	56.07%	56.19%	56.13%	67.42%	61.23%	64.18%
+Ngrams	2424	56.88%	56.78%	56.83%	68.37%	61.89%	64.97%
+Accounting Gazett.	2756	56.93%	56.86%	56.90%	67.97%	62.15%	64.93%
+Finance Gazett.	8261	59.86%	57.97%	58.90%	71.82%	66.51%	69.06%
+Window=1	8438	61.50%	58.98%	60.22%	72.71%	67.30%	69.90%
+POS Tag	8441	**61.50%**	**58.98%**	**60.22%**	**72.71%**	**67.30%**	**69.91%**
+SWN, Direction	8512	59.71%	57.63%	58.65%	71.96%	66.12%	68.92%
+Entropy	8438	59.82%	58.01%	58.90%	71.92%	67.04%	69.39%
+IDF	8417	59.81%	57.94%	58.86%	71.87%	66.87%	69.28%

TABLE A.8: Traditional & Modified Precision, Recall and F-Score with Classifier Combination, trained and tested using Ten Fold Cross-Validation, with a total of 1177 articles

APPENDIX B

Trading Simulation Results

B1 Long Strategy

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
3/01/95	12/01/95	NAB	10.339	10.419	967	76.97
6/01/95	19/01/95	BCD	1.150	1.150	8696	-1.74
9/01/95	17/01/95	ANZ	4.113	4.162	2431	119.61
13/01/95	23/01/95	BPC	3.310	3.316	3021	18.43
20/01/95	31/01/95	BEN	2.802	3.000	3569	705.95
27/01/95	10/02/95	BHP	19.057	18.222	525	-437.96
6/02/95	21/02/95	NAB	10.892	10.534	918	-328.64
15/02/95	23/02/95	ANZ	4.460	4.608	2242	331.14
17/02/95	8/03/95	BHP	18.001	18.187	556	103.36
28/02/95	8/03/95	NAB	10.898	10.699	918	-182.04
22/03/95	30/03/95	BHP	18.429	17.952	543	-258.85
23/03/95	31/03/95	ANZ	4.803	4.779	2082	-50.18
3/04/95	6/04/95	BHP	17.728	18.734	564	567.05
5/04/95	13/04/95	BEN	3.211	3.169	3114	-130.48
13/04/95	21/04/95	NAB	11.895	11.679	841	-181.32
12/04/95	4/05/95	BHP	19.231	20.223	520	516.05
12/05/95	22/05/95	AGL	4.378	4.315	2284	-145.03
15/05/95	23/05/95	ANZ	5.274	5.096	1896	-337.49
17/05/95	25/05/95	BIL	13.885	13.672	720	-153.50
19/05/95	29/05/95	NAB	11.761	11.761	850	0.17
22/05/95	30/05/95	BHP	17.656	17.765	566	61.41
30/05/95	8/06/95	ANZ	4.954	4.999	2018	90.00
5/06/95	13/06/95	BHP	17.434	16.894	574	-309.67
22/06/95	3/07/95	BIL	13.097	13.312	764	164.64
26/06/95	4/07/95	BEN	3.349	3.182	2986	-498.36
30/06/95	11/07/95	ANE	5.176	5.082	1932	-180.06
5/07/95	13/07/95	ANZ	5.091	5.231	1964	274.96
27/07/95	4/08/95	NAB	11.144	10.899	897	-219.68
29/06/95	11/08/95	BHP	17.779	19.420	562	922.35
16/08/95	24/08/95	ANZ	5.217	5.252	1917	68.05

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
17/08/95	25/08/95	BEN	3.386	3.440	2954	161.88
24/08/95	31/08/95	BHP	19.007	19.252	526	129.08
4/09/95	12/09/95	BEN	3.628	3.625	2756	-7.44
11/09/95	20/09/95	BHP	19.026	18.556	526	-246.96
21/09/95	26/09/95	BHP	18.443	18.151	542	-158.37
28/09/95	5/10/95	BHP	18.249	17.989	548	-142.75
16/10/95	24/10/95	BHP	18.254	17.559	548	-380.37
25/10/95	2/11/95	AGL	4.605	4.514	2172	-198.30
14/11/95	22/11/95	BEN	3.498	3.487	2858	-32.01
15/11/95	23/11/95	BIL	14.574	14.794	686	150.44
6/11/95	27/11/95	BHP	17.850	18.295	560	249.03
20/11/95	28/11/95	BPC	2.979	2.973	3357	-19.81
4/12/95	12/12/95	BHP	18.880	19.159	530	147.98
11/12/95	19/12/95	AGL	4.827	4.811	2072	-33.15
15/12/95	27/12/95	BHP	19.076	19.136	524	31.70
19/12/95	27/12/95	NAB	11.753	12.170	851	354.61
28/12/95	5/01/96	NAB	12.153	12.567	823	340.64
22/11/95	10/01/96	ANZ	5.899	6.398	1695	846.14
29/12/95	15/01/96	BHP	19.048	18.788	525	-136.45
16/01/96	17/01/96	BHP	18.841	18.966	531	65.90
11/01/96	19/01/96	BPC	3.005	2.913	3328	-306.51
22/01/96	23/01/96	BHP	18.724	18.703	534	-11.37
29/01/96	6/02/96	NAB	12.725	12.794	786	54.31
8/02/96	16/02/96	ADB	4.126	4.149	2424	55.51
9/02/96	21/02/96	BHP	18.873	18.471	530	-212.95
22/02/96	1/03/96	ASL	0.931	0.929	10743	-18.26
23/02/96	4/03/96	BHP	18.789	19.323	532	284.30
5/03/96	8/03/96	BHP	19.112	18.678	523	-226.98
11/03/96	27/03/96	BHP	18.251	18.297	548	24.82
25/03/96	2/04/96	BEN	4.115	3.984	2430	-316.87
28/03/96	10/04/96	BHP	18.180	18.293	550	61.99
11/04/96	17/04/96	AGM	0.085	0.111	117786	3038.88
12/04/96	22/04/96	BHP	19.007	19.279	526	143.28
26/04/96	6/05/96	BHP	19.796	19.679	505	-58.88
30/04/96	13/05/96	ANZ	6.106	5.966	1638	-230.47
8/05/96	20/05/96	BHP	19.973	19.037	501	-468.99
17/05/96	27/05/96	NAB	11.644	11.872	859	195.77
23/05/96	27/05/96	AVJ	0.532	0.633	18786	1887.99
24/05/96	3/06/96	ANZ	6.019	5.603	1661	-691.47
31/05/96	11/06/96	BHP	18.908	18.677	529	-122.62
14/06/96	19/06/96	BHP	18.599	18.156	538	-238.12

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
24/06/96	2/07/96	AGL	5.220	5.258	1916	72.23
2/07/96	12/07/96	ANZ	6.178	5.997	1619	-291.91
9/07/96	17/07/96	ALL	3.257	3.271	3070	43.59
17/07/96	25/07/96	NAB	11.244	11.673	889	381.38
12/08/96	13/08/96	BGF	0.065	0.061	153374	-705.52
19/08/96	27/08/96	ANZ	6.622	6.607	1510	-22.20
27/08/96	4/09/96	BEN	3.741	3.832	2673	244.85
30/08/96	5/09/96	BHP	17.264	17.204	579	-34.86
6/09/96	18/09/96	AVJ	0.590	0.599	16949	145.76
20/09/96	23/09/96	BHP	16.012	15.812	625	-124.81
17/09/96	25/09/96	BGF	0.066	0.059	151057	-1072.50
19/09/96	27/09/96	BEN	3.891	3.951	2570	152.92
1/10/96	2/10/96	ANE	3.475	3.448	2878	-76.55
4/10/96	14/10/96	ANZ	7.261	7.339	1377	107.82
17/10/96	18/10/96	ANZ	7.298	7.295	1370	-4.38
10/10/96	18/10/96	BEN	3.964	4.089	2523	314.87
28/10/96	5/11/96	BEN	4.173	4.190	2396	40.49
31/10/96	8/11/96	ANZ	7.341	7.437	1362	129.93
5/11/96	13/11/96	NAB	13.883	13.712	720	-123.19
7/11/96	15/11/96	BHP	16.560	17.445	604	534.66
11/11/96	19/11/96	BIL	22.173	21.834	451	-152.62
15/11/96	25/11/96	NAB	13.911	14.516	719	435.14
21/11/96	29/11/96	ABC	1.700	1.660	5881	-234.65
27/11/96	5/12/96	BIL	21.742	21.663	460	-36.39
5/12/96	13/12/96	NAB	14.938	14.661	669	-185.38
10/12/96	16/12/96	ANZ	7.768	7.532	1287	-304.50
12/12/96	18/12/96	BHP	17.428	16.987	574	-252.73
17/12/96	27/12/96	NAB	14.237	14.556	702	224.01
20/12/96	30/12/96	BGF	0.054	0.055	183824	110.29
2/01/97	10/01/97	BEN	4.178	4.194	2393	37.33
7/01/97	16/01/97	NAB	14.762	14.432	677	-223.07
9/01/97	17/01/97	AGL	7.042	6.894	1420	-210.02
31/01/97	10/02/97	ADB	4.596	4.755	2176	346.20
24/01/97	11/02/97	NAB	15.775	15.848	634	46.73
5/02/97	13/02/97	BHP	17.632	17.966	567	189.26
12/02/97	20/02/97	AXI	0.430	0.446	23256	365.12
18/02/97	26/02/97	BKR	1.314	1.300	7609	-108.81
24/02/97	4/03/97	BHP	17.787	17.012	562	-435.94
25/02/97	5/03/97	AVJ	0.628	0.619	15921	-146.47
5/03/97	13/03/97	NAB	16.055	16.006	623	-30.46
6/03/97	14/03/97	ARG	2.805	2.848	3565	152.94

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
11/03/97	19/03/97	BPC	2.074	2.109	4823	172.66
18/03/97	26/03/97	ALL	2.882	2.669	3470	-738.76
19/03/97	27/03/97	NAB	15.788	16.155	633	232.50
27/03/97	4/04/97	ANZ	8.109	8.015	1233	-115.90
3/04/97	11/04/97	BHP	16.806	16.837	595	18.45
15/04/97	1/05/97	AXI	0.450	0.460	22222	222.22
8/05/97	16/05/97	NAB	18.069	17.914	553	-85.99
14/05/97	22/05/97	BIL	23.721	23.747	422	10.72
29/05/97	6/06/97	AGL	7.765	7.873	1288	139.49
6/06/97	16/06/97	BHP	18.706	18.907	535	107.48
11/06/97	19/06/97	NAB	18.923	18.757	528	-87.70
17/06/97	25/06/97	ALZ	1.534	1.502	6521	-203.46
28/05/97	11/07/97	ANZ	8.657	10.134	1155	1705.70
9/07/97	22/07/97	BHP	18.969	17.934	527	-545.71
18/07/97	28/07/97	ANZ	10.036	10.509	996	471.61
31/07/97	6/08/97	BHP	18.255	17.970	548	-156.02
21/08/97	1/09/97	BHP	17.360	17.058	576	-173.89
29/08/97	8/09/97	AGL	8.537	8.937	1171	468.63
2/09/97	10/09/97	AEX	0.033	0.032	304878	-243.90
5/09/97	15/09/97	ANZ	9.873	10.423	1013	557.25
16/09/97	17/09/97	BHP	16.131	16.251	620	74.83
19/09/97	22/09/97	BHP	15.798	15.665	633	-84.63
2/10/97	10/10/97	BHP	16.322	15.768	613	-339.72
27/10/97	4/11/97	ANZ	9.633	10.415	1038	811.20
5/11/97	19/11/97	ALZ	1.330	1.281	7519	-369.93
19/11/97	1/12/97	BHP	13.179	13.749	759	432.93
25/11/97	3/12/97	BEN	3.435	3.441	2911	16.01
5/12/97	16/12/97	BPC	0.248	0.287	40371	1598.69
22/12/97	23/12/97	BPC	0.271	0.269	36955	-77.61
15/12/97	23/12/97	ANZ	9.985	9.774	1001	-211.21
18/12/97	30/12/97	BKR	1.026	1.030	9751	43.88
5/01/98	13/01/98	BKL	4.341	4.650	2304	712.17
16/01/98	29/01/98	AUO	0.412	0.440	24295	682.69
23/01/98	2/02/98	ANZ	9.651	9.975	1036	335.46
4/02/98	19/02/98	ANC	2.625	2.750	3810	476.25
17/02/98	25/02/98	ANZ	9.487	9.848	1054	379.97
2/03/98	10/03/98	BHP	14.541	15.876	688	918.34
9/03/98	17/03/98	ARQ	0.430	0.426	23256	-95.35
13/03/98	19/03/98	BPC	0.252	0.242	39761	-397.61
26/02/98	20/03/98	ADZ	3.030	3.090	3301	198.06
25/03/98	2/04/98	ANC	3.517	3.592	2843	214.93

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
6/04/98	14/04/98	BIL	31.259	33.008	320	559.87
28/04/98	6/05/98	BEN	3.737	3.748	2676	29.70
30/04/98	8/05/98	ANZ	10.723	11.072	933	325.62
1/05/98	11/05/98	BHP	15.053	14.352	664	-465.46
14/05/98	25/05/98	BHP	14.127	13.689	708	-310.46
1/06/98	12/06/98	BHP	13.740	13.141	728	-435.56
18/06/98	26/06/98	ALZ	1.004	1.000	9959	-40.83
19/06/98	29/06/98	ADB	5.232	5.201	1911	-58.09
30/06/98	1/07/98	AXI	0.094	0.150	106270	5940.49
23/06/98	3/07/98	BHP	13.725	14.043	729	232.04
1/07/98	10/07/98	BEN	3.857	3.968	2593	288.34
2/07/98	10/07/98	AMP	20.173	19.757	496	-206.39
6/07/98	14/07/98	BGF	0.024	0.024	416667	83.33
16/07/98	24/07/98	ANZ	11.603	10.901	862	-604.87
13/07/98	27/07/98	BHP	14.014	13.886	714	-91.61
15/07/98	3/08/98	AMP	19.489	21.333	513	945.87
3/08/98	13/08/98	BHP	13.485	12.874	742	-452.99
10/08/98	20/08/98	BDS	3.200	3.000	3125	-625.00
22/06/98	21/08/98	ALL	3.549	4.097	2818	1544.83
24/08/98	1/09/98	ADA	0.701	0.650	14257	-732.81
21/08/98	4/09/98	AMP	22.205	21.153	450	-473.76
28/08/98	7/09/98	BIL	34.239	33.967	292	-79.42
31/08/98	9/09/98	BHP	12.157	12.800	823	529.77
1/09/98	10/09/98	AIZ	1.658	1.600	6030	-352.15
11/09/98	14/09/98	BHP	12.259	12.471	816	172.83
17/09/98	21/09/98	BHP	12.551	12.371	797	-143.70
2/09/98	24/09/98	ALL	4.043	4.796	2473	1861.43
24/09/98	28/09/98	ANZ	9.056	8.985	1104	-78.72
1/10/98	9/10/98	AMP	20.020	18.558	500	-731.05
12/10/98	14/10/98	BHP	12.331	12.248	811	-67.80
13/10/98	21/10/98	ANZ	9.053	8.936	1105	-129.29
16/10/98	26/10/98	BHP	12.550	13.207	797	523.55
2/11/98	10/11/98	ADB	5.776	5.778	1731	2.25
2/11/98	10/11/98	BHP	13.513	13.420	740	-68.82
5/11/98	13/11/98	AMP	19.308	19.829	518	270.09
17/11/98	25/11/98	ADZ	2.772	2.790	3608	67.47
18/11/98	26/11/98	AVJ	0.442	0.440	22650	-38.51
13/11/98	2/12/98	BEN	4.644	5.367	2153	1556.19
4/12/98	9/12/98	ADZ	3.026	3.014	3305	-38.34
1/12/98	9/12/98	ASX	8.714	7.493	1148	-1401.71
30/11/98	10/12/98	BHP	12.673	11.877	789	-627.81

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
3/12/98	11/12/98	ALZ	1.065	1.067	9387	16.90
7/12/98	15/12/98	BEN	5.206	5.387	1921	347.70
22/12/98	30/12/98	AMP	19.999	20.271	500	135.85
6/01/99	14/01/99	BEN	6.280	5.744	1592	-853.31
13/01/99	18/01/99	BHP	12.775	12.123	783	-511.06
8/01/99	22/01/99	ANZ	10.972	10.306	911	-607.27
1/02/99	9/02/99	AGL	11.347	11.079	881	-235.93
11/02/99	15/02/99	BHP	12.111	11.910	826	-166.11
8/02/99	16/02/99	AMP	19.169	18.686	522	-251.92
16/02/99	24/02/99	ALZ	1.311	1.347	7628	277.66
24/02/99	4/03/99	ALL	9.730	9.704	1028	-27.14
22/02/99	9/03/99	ADG	1.951	2.000	5126	252.71
18/03/99	26/03/99	ASX	15.393	14.220	650	-762.58
29/03/99	6/04/99	ASX	13.812	14.622	724	586.87
7/04/99	15/04/99	AMP	17.549	18.191	570	365.94
13/04/99	19/04/99	BHP	14.213	16.788	704	1813.01
6/04/99	20/04/99	ARQ	0.237	0.230	42248	-283.06
12/04/99	20/04/99	BCD	1.215	1.176	8230	-322.62
20/04/99	22/04/99	BHP	16.544	15.855	604	-416.04
14/04/99	22/04/99	BIL	43.362	44.460	231	253.71
16/04/99	27/04/99	ASX	14.064	14.811	711	531.19
3/05/99	10/05/99	AMP	17.517	17.559	571	23.81
7/05/99	24/05/99	BHP	17.511	16.958	571	-315.99
20/05/99	4/06/99	ADA	1.445	1.481	6919	243.55
21/05/99	8/06/99	BIL	40.852	41.034	245	44.66
26/05/99	15/06/99	AVJ	0.492	0.572	20313	1612.85
4/06/99	15/06/99	AMP	16.627	15.608	601	-612.54
11/06/99	21/06/99	BHP	17.488	17.666	572	101.82
15/06/99	23/06/99	ASX	10.281	10.075	973	-199.66
21/06/99	29/06/99	AMP	16.280	16.719	614	269.36
25/06/99	5/07/99	ASX	10.006	10.779	999	771.93
29/06/99	7/07/99	BPC	0.294	0.308	34060	497.28
9/07/99	20/07/99	BHP	18.716	18.519	534	-105.20
15/07/99	23/07/99	ANZ	11.409	11.378	877	-27.45
16/07/99	26/07/99	BIL	40.549	39.971	247	-142.77
20/07/99	28/07/99	AMP	17.171	17.301	582	75.95
23/07/99	2/08/99	ALZ	1.463	1.431	6834	-221.42
28/07/99	5/08/99	AGL	9.964	10.124	1004	160.54
29/07/99	6/08/99	BHP	17.384	18.203	575	471.16
30/07/99	9/08/99	AMP	17.440	16.448	573	-568.65
9/08/99	17/08/99	AMP	16.448	15.839	608	-369.97

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
20/08/99	31/08/99	ADG	2.219	2.270	4506	228.45
26/08/99	1/09/99	ASX	11.664	11.657	857	-5.91
24/08/99	6/09/99	AMP	15.875	15.737	630	-86.63
6/09/99	14/09/99	BKW	41.748	42.000	240	60.60
13/09/99	15/09/99	ADA	1.796	2.312	5568	2871.42
20/09/99	21/09/99	BHP	18.397	18.203	544	-105.26
14/09/99	22/09/99	AUN	5.102	4.757	1960	-676.00
22/09/99	27/09/99	BGN	1.154	1.543	8663	3362.98
5/10/99	13/10/99	ANZ	10.195	10.372	981	173.74
1/10/99	20/10/99	AMP	14.381	14.763	695	265.91
20/10/99	28/10/99	AGL	8.847	8.689	1130	-179.22
26/10/99	3/11/99	ASX	9.239	9.405	1082	179.72
8/11/99	16/11/99	BCD	1.073	1.050	9319	-211.54
9/11/99	17/11/99	ASX	9.392	9.605	1065	227.27
4/11/99	22/11/99	BHP	16.495	18.471	606	1197.21
17/11/99	25/11/99	AUN	5.612	5.282	1782	-588.06
24/11/99	2/12/99	ALZ	1.470	1.520	6805	342.97
29/11/99	8/12/99	BHP	17.698	17.469	565	-129.50
2/12/99	10/12/99	BKL	5.529	5.249	1809	-506.52
7/12/99	15/12/99	AGL	8.731	8.968	1145	270.68
15/12/99	23/12/99	ANZ	11.504	11.020	869	-420.77
22/12/99	23/12/99	BNO	0.829	1.045	12070	2610.74
20/12/99	29/12/99	AGL	9.201	8.975	1087	-245.12
20/12/99	29/12/99	BHP	19.947	19.934	501	-6.51
29/12/99	30/12/99	ASC	1.859	1.453	5380	-2180.51
23/12/99	4/01/00	AMP	16.726	16.536	598	-113.62
6/01/00	11/01/00	BHP	20.452	20.967	489	252.23
4/01/00	12/01/00	ANZ	10.822	10.312	924	-470.78
21/12/99	18/01/00	ASX	11.071	13.100	903	1832.10
17/01/00	25/01/00	AMP	16.694	15.689	599	-601.88
18/01/00	1/02/00	ANZ	11.052	10.413	905	-577.57
31/01/00	7/02/00	AGM	0.181	0.153	55188	-1567.34
31/01/00	8/02/00	BHP	18.604	18.928	538	174.26
3/02/00	14/02/00	AMP	14.869	14.514	673	-239.25
25/01/00	16/02/00	AUN	7.059	8.128	1417	1514.35
10/02/00	18/02/00	BHP	17.686	16.504	565	-667.49
11/02/00	21/02/00	ANZ	10.421	10.065	960	-340.99
14/02/00	23/02/00	ASX	13.704	14.334	730	460.05
1/03/00	9/03/00	BHP	16.237	16.759	616	321.24
3/03/00	13/03/00	ASX	14.918	14.934	670	10.45
29/02/00	14/03/00	ADZ	2.147	1.908	4658	-1111.86

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
6/03/00	14/03/00	ANZ	10.210	10.379	979	166.04
21/03/00	23/03/00	AET	2.067	1.432	4838	-3072.13
29/03/00	6/04/00	ANZ	10.689	11.001	936	292.31
3/04/00	11/04/00	ADB	4.684	4.696	2135	25.19
20/03/00	12/04/00	BHP	17.739	17.969	564	129.49
14/04/00	26/04/00	BPC	0.404	0.405	24740	27.21
28/04/00	5/05/00	ANZ	11.968	11.564	836	-337.74
27/04/00	5/05/00	BLD	2.000	2.055	4999	271.95
1/05/00	9/05/00	ASX	12.452	11.410	803	-837.21
10/05/00	18/05/00	BIL	48.527	46.207	206	-478.00
11/05/00	19/05/00	ASB	2.242	2.150	4460	-410.32
8/05/00	22/05/00	BHP	17.765	17.048	563	-403.73
18/05/00	26/05/00	AXA	2.562	2.562	3903	-3.12
15/05/00	29/05/00	ASX	10.705	9.749	934	-892.81
24/05/00	29/05/00	BHP	17.425	17.492	574	38.46
29/05/00	6/06/00	AXA	2.576	2.597	3882	81.91
1/06/00	9/06/00	BHP	17.633	17.729	567	54.66
7/06/00	15/06/00	ASX	10.386	10.111	963	-265.50
9/06/00	19/06/00	ALZ	1.338	1.311	7476	-197.37
20/06/00	29/06/00	BIL	49.874	50.958	201	217.96
23/06/00	3/07/00	AIZ	1.697	1.717	5893	115.50
27/06/00	5/07/00	BHP	19.537	19.090	512	-229.02
30/06/00	10/07/00	ANZ	12.777	12.808	783	23.96
5/07/00	17/07/00	BBB	0.736	0.718	13583	-248.57
11/07/00	19/07/00	BHP	19.246	18.780	520	-242.22
12/07/00	28/07/00	AUO	0.125	0.147	80000	1744.00
20/07/00	28/07/00	BPC	0.461	0.474	21706	282.18
26/07/00	3/08/00	ARG	3.433	3.412	2913	-58.55
28/07/00	4/08/00	BHP	18.289	18.286	547	-1.91
27/07/00	7/08/00	ALZ	1.476	1.468	6773	-58.93
31/07/00	8/08/00	AMP	17.323	17.315	577	-4.62
9/08/00	17/08/00	ASX	11.096	11.432	901	301.93
2/08/00	18/08/00	AXA	2.531	2.620	3951	349.66
21/08/00	28/08/00	AMP	17.732	17.998	564	149.91
24/08/00	1/09/00	ASX	12.191	12.242	820	42.07
29/08/00	14/09/00	BBB	0.676	0.643	14799	-479.49
6/09/00	14/09/00	ANZ	13.102	13.163	763	46.70
5/09/00	19/09/00	AUN	5.142	4.366	1945	-1508.74
11/09/00	22/09/00	BHP	20.075	20.192	498	58.32
3/10/00	11/10/00	BHP	19.469	19.709	514	123.31
29/09/00	12/10/00	ANZ	13.234	13.839	756	457.23

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
12/10/00	20/10/00	BDL	2.201	2.019	4543	-830.01
17/10/00	23/10/00	BBB	0.501	0.593	19952	1837.58
10/10/00	26/10/00	BEN	5.198	5.155	1924	-84.27
23/10/00	31/10/00	ADB	5.036	5.004	1986	-62.96
25/10/00	2/11/00	AGL	11.739	11.619	852	-102.41
9/10/00	16/11/00	ASX	11.012	12.802	908	1625.14
20/11/00	28/11/00	AUN	2.948	2.489	3392	-1558.62
20/11/00	28/11/00	AUN	2.948	2.489	3392	-1558.62
20/11/00	5/12/00	BDL	2.607	2.472	3836	-517.48
27/11/00	7/12/00	AXA	2.803	2.821	3568	64.22
27/12/00	4/01/01	ANZ	14.840	14.690	674	-101.30
28/12/00	8/01/01	BHP	19.190	19.708	521	269.83
11/01/01	19/01/01	ALZ	1.250	1.208	8000	-337.60
13/11/00	23/01/01	AMP	17.991	18.685	556	385.86
15/01/01	23/01/01	AET	2.400	2.320	4167	-333.36
18/01/01	29/01/01	BEN	5.934	6.165	1685	390.08
16/01/01	1/02/01	BIL	42.991	49.567	233	1532.14
29/01/01	6/02/01	ANZ	14.865	14.980	673	77.13
31/01/01	8/02/01	BBG	4.359	4.311	2294	-109.88
6/02/01	16/02/01	AXA	2.675	2.646	3739	-105.44
20/02/01	28/02/01	BEN	6.016	6.228	1662	352.51
1/03/01	9/03/01	BBB	0.250	0.231	40000	-760.00
9/02/01	19/03/01	BHP	18.689	21.612	535	1563.75
16/03/01	21/03/01	AUN	1.071	0.883	9336	-1759.84
19/03/01	26/03/01	BHP	21.612	19.978	463	-756.73
26/03/01	30/03/01	BBB	0.201	0.231	49776	1503.24
3/04/01	11/04/01	AGL	10.843	11.039	922	180.99
5/04/01	17/04/01	ANZ	14.451	13.623	692	-572.98
9/04/01	17/04/01	BOQ	6.626	6.363	1509	-396.57
12/04/01	19/04/01	BHP	21.379	21.700	468	150.23
27/04/01	1/05/01	ANZ	14.017	13.963	713	-38.29
7/05/01	15/05/01	AXA	2.842	2.780	3519	-218.18
10/05/01	18/05/01	AGL	10.714	10.474	933	-223.36
14/05/01	18/05/01	AVJ	0.729	0.613	13710	-1598.59
15/05/01	23/05/01	AMP	20.049	20.444	499	196.76
21/05/01	24/05/01	BHP	23.202	23.205	431	1.08
16/05/01	24/05/01	ANZ	14.801	15.053	676	169.95
23/05/01	31/05/01	AIZ	1.173	1.173	8524	0.85
4/06/01	15/06/01	AXA	2.831	2.845	3533	51.58
6/06/01	15/06/01	ADB	5.471	5.551	1828	145.51
7/06/01	15/06/01	AMP	20.495	21.254	488	370.78

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
18/06/01	26/06/01	BBB	0.197	0.181	50736	-832.07
27/06/01	5/07/01	AMP	21.184	20.336	472	-400.07
3/07/01	11/07/01	ANM	2.090	2.080	4784	-49.28
6/07/01	16/07/01	ABS	3.507	4.140	2852	1806.46
29/06/01	23/07/01	BLD	2.860	3.300	3497	1539.03
2/07/01	24/07/01	ALZ	1.570	1.603	6369	208.27
23/07/01	31/07/01	AIZ	1.133	1.099	8830	-297.57
7/08/01	17/08/01	AMP	19.093	19.012	524	-42.81
17/08/01	27/08/01	ALZ	1.646	1.620	6075	-157.95
15/08/01	31/08/01	ALL	7.155	7.216	1398	85.56
24/08/01	3/09/01	BLD	3.393	3.483	2948	265.32
29/08/01	6/09/01	BKW	7.621	7.263	1312	-469.43
11/09/01	14/09/01	BBB	0.100	0.084	100000	-1650.00
10/09/01	18/09/01	AHD	1.887	1.707	5299	-952.76
1/08/01	19/09/01	ADB	6.132	6.410	1631	453.09
21/09/01	1/10/01	AMP	17.305	18.334	578	594.82
1/10/01	9/10/01	BEN	6.093	6.085	1641	-12.47
5/10/01	15/10/01	ANZ	16.871	17.099	593	135.20
12/10/01	22/10/01	BHP	9.489	9.072	1054	-439.62
18/10/01	24/10/01	BBB	0.085	0.119	117647	3988.23
2/11/01	6/11/01	BHP	8.984	8.990	1113	6.57
29/10/01	6/11/01	ANZ	17.832	17.936	561	58.62
5/11/01	9/11/01	AWB	4.003	3.920	2498	-205.59
14/11/01	15/11/01	AUN	0.293	0.345	34095	1772.94
4/12/01	10/12/01	AUN	0.315	0.308	31716	-244.21
7/11/01	10/12/01	BHP	9.016	10.346	1109	1474.75
3/12/01	11/12/01	AMP	19.313	18.582	518	-378.40
26/10/01	17/12/01	BBG	7.682	7.788	1302	138.27
12/12/01	20/12/01	ASX	11.611	11.412	861	-171.25
14/12/01	2/01/02	AXA	2.750	2.890	3636	508.68
19/12/01	2/01/02	BCA	6.048	6.200	1653	251.09
19/12/01	2/01/02	AGI	1.127	1.164	8876	331.07
7/01/02	15/01/02	ALZ	1.820	1.828	5495	43.96
17/01/02	18/01/02	AMP	18.741	18.607	534	-71.72
10/01/02	18/01/02	ARG	4.457	4.445	2244	-28.50
15/01/02	23/01/02	AUN	0.334	0.333	29976	-8.99
29/01/02	5/02/02	BHP	11.227	11.731	891	448.53
25/01/02	8/02/02	ALZ	1.739	1.691	5749	-276.53
7/02/02	15/02/02	AMP	18.509	18.477	540	-17.50
11/02/02	18/02/02	BHP	11.804	11.688	847	-98.17
4/02/02	18/02/02	BDS	1.542	1.852	6486	2010.66

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
8/02/02	18/02/02	AEO	2.158	2.209	4634	237.26
11/02/02	19/02/02	AUN	0.344	0.324	29036	-606.85
12/02/02	20/02/02	ARG	4.630	4.631	2160	2.38
14/02/02	22/02/02	BLD	4.001	4.048	2500	118.50
19/02/02	1/03/02	BEN	7.683	7.374	1302	-402.19
22/02/02	4/03/02	AXA	2.882	3.045	3470	565.61
26/02/02	5/03/02	AUN	0.369	0.298	27130	-1912.67
7/03/02	15/03/02	BCA	5.323	5.217	1879	-197.67
8/03/02	18/03/02	BOQ	7.039	6.968	1421	-101.46
18/03/02	27/03/02	APE	4.660	4.792	2146	282.20
25/03/02	2/04/02	BIL	9.870	9.605	1013	-268.45
5/04/02	15/04/02	ANZ	17.828	18.191	561	203.25
8/04/02	16/04/02	BOQ	7.139	7.227	1401	123.29
12/04/02	22/04/02	AGL	9.322	9.409	1073	93.24
3/04/02	23/04/02	BHP	11.739	11.397	852	-290.96
17/04/02	26/04/02	ANZ	18.346	19.414	545	582.33
1/05/02	3/05/02	BHP	10.809	10.839	925	27.29
8/04/02	7/05/02	ASB	1.526	1.548	6555	146.18
29/04/02	7/05/02	ANZ	19.257	19.080	519	-91.71
6/05/02	13/05/02	AUN	0.153	0.179	65317	1711.31
9/05/02	17/05/02	BBB	0.158	0.155	63371	-177.44
13/05/02	21/05/02	BHP	11.152	11.249	897	86.56
14/05/02	23/05/02	AUN	0.209	0.213	47755	157.59
6/05/02	27/05/02	AIZ	0.431	0.497	23207	1536.30
17/05/02	27/05/02	BIL	9.480	9.519	1055	40.41
24/05/02	3/06/02	AGL	9.688	9.838	1032	155.32
27/05/02	4/06/02	ALL	5.885	5.923	1699	65.24
31/05/02	6/06/02	AXA	3.080	3.105	3247	81.18
11/06/02	17/06/02	BHP	10.278	10.325	973	45.15
21/06/02	26/06/02	BHP	10.250	10.076	976	-169.73
2/07/02	12/07/02	BHP	10.144	10.036	986	-106.78
5/07/02	15/07/02	ANZ	18.927	19.220	528	154.86
17/07/02	22/07/02	ASX	12.959	12.499	772	-354.97
15/07/02	22/07/02	BHP	9.927	9.759	1007	-169.28
15/07/02	23/07/02	ANZ	19.220	18.490	520	-379.55
18/07/02	26/07/02	BOQ	7.364	7.005	1358	-488.34
23/07/02	29/07/02	BHP	9.744	9.161	1026	-598.47
12/07/02	1/08/02	AXI	0.023	0.035	438596	5350.87
24/07/02	2/08/02	ARQ	0.386	0.393	25920	181.44
29/07/02	6/08/02	BEN	6.815	7.076	1467	382.45
2/08/02	8/08/02	BHP	9.010	8.960	1110	-55.83

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
9/08/02	19/08/02	BBG	7.909	8.314	1264	511.54
12/08/02	20/08/02	BOQ	7.042	7.165	1420	174.38
13/08/02	21/08/02	BHP	9.030	9.298	1107	296.23
19/08/02	27/08/02	ASX	12.429	12.700	805	218.32
20/08/02	28/08/02	AMP	14.474	14.287	691	-129.56
26/08/02	3/09/02	ABS	13.020	15.303	768	1753.50
29/08/02	10/09/02	ASX	12.589	12.406	794	-145.86
2/09/02	10/09/02	APE	5.158	5.100	1939	-113.04
28/08/02	12/09/02	AXA	2.610	2.582	3832	-104.23
6/09/02	16/09/02	AUN	0.183	0.171	54526	-665.22
13/09/02	23/09/02	BBG	7.948	7.689	1258	-325.57
16/09/02	24/09/02	BHP	9.431	8.864	1060	-601.02
17/09/02	25/09/02	ASL	0.430	0.422	23240	-185.92
26/09/02	4/10/02	AMP	12.028	11.793	831	-195.70
30/09/02	8/10/02	AUN	0.141	0.148	70721	445.54
30/09/02	11/10/02	BHP	9.123	9.147	1096	26.19
14/10/02	23/10/02	ANN	6.674	6.819	1498	216.91
15/10/02	23/10/02	ALZ	1.303	1.366	7675	481.22
14/10/02	25/10/02	BHP	9.471	9.690	1056	230.95
17/10/02	25/10/02	BIL	6.826	6.679	1465	-214.92
18/10/02	29/10/02	BOQ	7.139	6.949	1401	-266.05
24/10/02	1/11/02	AMP	13.029	12.440	768	-452.35
30/10/02	7/11/02	AIZ	0.462	0.430	21654	-684.27
31/10/02	13/11/02	BHP	9.582	9.436	1044	-151.69
7/11/02	15/11/02	AMP	12.232	12.373	818	115.17
8/11/02	18/11/02	AEO	1.422	1.485	7035	448.13
11/11/02	19/11/02	AWB	3.478	3.435	2875	-124.78
21/11/02	29/11/02	ASX	11.327	11.546	883	193.38
15/11/02	3/12/02	BCA	1.890	1.802	5292	-465.17
13/01/03	21/01/03	ADZ	1.899	1.839	5267	-314.44
14/01/03	22/01/03	BBG	7.408	7.091	1350	-428.09
16/01/03	24/01/03	AGI	0.654	0.654	15300	7.65
4/02/03	12/02/03	AMP	9.335	8.484	1071	-910.67
5/02/03	13/02/03	ALZ	1.425	1.439	7020	98.28
10/02/03	3/03/03	BLD	4.312	4.469	2319	365.47
21/02/03	3/03/03	AXA	2.282	2.205	4382	-339.61
3/03/03	6/03/03	BHP	9.262	8.976	1080	-309.42
28/02/03	10/03/03	AIZ	0.480	0.460	20838	-414.68
3/03/03	11/03/03	AHO	0.380	0.370	26316	-263.16
5/03/03	13/03/03	ABS	2.404	2.220	4161	-763.54
8/04/03	16/04/03	AQA	0.423	0.435	23629	278.82

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
11/03/03	8/05/03	BHP	8.604	8.530	1162	-86.69
28/03/03	14/05/03	AWB	3.553	3.741	2815	530.35
6/05/03	14/05/03	ANZ	18.341	18.425	545	45.56
22/05/03	30/05/03	AWB	3.778	3.888	2647	291.70
3/06/03	10/06/03	AMP	5.002	4.983	1999	-38.38
2/06/03	10/06/03	ANZ	18.698	18.864	535	88.92
6/06/03	16/06/03	BBB	0.105	0.110	95147	466.22
12/06/03	23/06/03	AMP	4.925	5.164	2030	483.75
16/06/03	24/06/03	ALN	5.665	5.880	1765	379.65
19/06/03	27/06/03	ALZ	1.631	1.613	6132	-107.92
23/06/03	30/06/03	AGI	0.426	0.377	23491	-1148.71
18/06/03	1/07/03	ALL	1.308	1.326	7646	140.69
27/06/03	7/07/03	ALN	5.633	5.615	1775	-32.13
2/07/03	10/07/03	ASX	12.817	12.979	780	126.59
9/07/03	17/07/03	BHP	8.961	9.246	1116	317.95
11/07/03	22/07/03	ALN	5.733	5.801	1744	119.99
18/07/03	23/07/03	BHP	9.334	9.366	1071	34.06
15/07/03	23/07/03	ANE	2.791	2.840	3583	175.57
25/07/03	4/08/03	ANZ	18.052	18.145	554	51.80
28/07/03	5/08/03	BHP	9.988	9.997	1001	9.21
29/07/03	6/08/03	ALZ	1.681	1.742	5950	361.17
15/08/03	25/08/03	AXA	2.542	2.689	3934	580.27
18/08/03	26/08/03	BOQ	8.843	8.757	1131	-97.15
19/08/03	28/08/03	AMP	4.784	6.287	2091	3142.77
25/08/03	4/09/03	APE	6.840	6.766	1462	-108.33
25/08/03	2/09/03	BHP	10.622	11.229	941	571.66
5/09/03	15/09/03	ADB	8.709	8.465	1148	-279.88
8/09/03	16/09/03	BBG	7.227	7.418	1384	263.24
8/09/03	16/09/03	AWB	3.818	3.923	2619	274.21
11/09/03	19/09/03	ALN	6.101	6.459	1639	587.58
23/09/03	1/10/03	AIZ	0.446	0.447	22427	15.70
15/09/03	6/10/03	ALL	1.890	1.946	5291	296.30
7/10/03	14/10/03	AMP	6.613	6.969	1512	538.42
1/10/03	16/10/03	BHP	10.344	12.120	967	1716.72
17/10/03	27/10/03	BOQ	9.525	9.725	1050	210.11
22/10/03	30/10/03	ASX	15.442	15.955	648	332.75
24/10/03	31/10/03	BHP	11.391	11.728	878	295.89
28/10/03	5/11/03	ALZ	1.711	1.736	5846	147.32
3/11/03	7/11/03	BBB	0.231	0.279	43215	2052.71
5/11/03	13/11/03	AXA	2.831	2.732	3532	-351.79
18/11/03	26/11/03	AGI	0.547	0.539	18275	-148.03

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
26/11/03	4/12/03	ALZ	1.648	1.637	6066	-70.97
1/12/03	9/12/03	BHP	11.112	11.549	900	393.12
1/12/03	9/12/03	AMP	5.918	5.739	1690	-303.19
2/12/03	10/12/03	AWB	4.282	4.260	2335	-50.44
9/12/03	17/12/03	BLT	1.434	1.250	6974	-1284.61
12/12/03	22/12/03	ALL	1.721	1.672	5811	-285.90
17/12/03	29/12/03	BLD	5.041	5.053	1984	23.61
19/12/03	29/12/03	BBB	0.306	0.282	32637	-799.61
29/12/03	6/01/04	AGL	11.200	11.152	893	-42.51
6/01/04	14/01/04	BBG	6.963	6.879	1436	-121.34
14/01/04	28/01/04	BHP	11.779	11.488	849	-247.40
30/01/04	9/02/04	BHP	11.296	11.556	885	229.83
2/02/04	10/02/04	ALZ	1.652	1.659	6053	39.34
2/02/04	10/02/04	ALN	6.652	6.893	1503	363.58
10/02/04	18/02/04	ARG	5.170	5.315	1934	280.04
12/02/04	20/02/04	BDS	2.870	2.864	3484	-21.60
6/02/04	24/02/04	ANZ	17.683	17.823	566	78.79
18/02/04	26/02/04	AGI	0.540	0.532	18508	-151.77
20/02/04	1/03/04	BHP	12.423	12.399	805	-19.64
23/02/04	2/03/04	ALZ	1.702	1.764	5876	367.25
27/02/04	8/03/04	ALL	2.403	2.886	4162	2013.99
3/03/04	11/03/04	BOQ	10.982	11.408	911	387.45
4/03/04	12/03/04	ASB	1.006	0.987	9942	-182.93
8/03/04	16/03/04	AWB	4.892	4.837	2044	-112.22
1/03/04	23/03/04	ALN	7.186	6.012	1392	-1634.76
18/03/04	29/03/04	ADB	8.527	8.400	1173	-148.97
24/03/04	1/04/04	BHP	12.026	12.350	832	269.57
25/03/04	2/04/04	AIZ	0.338	0.358	29621	592.42
17/02/04	6/04/04	BEN	9.369	10.870	1067	1601.67
31/03/04	8/04/04	AXA	3.200	3.238	3125	116.56
2/03/04	16/04/04	ANZ	18.114	18.611	552	274.12
8/04/04	16/04/04	BOQ	11.751	11.090	851	-562.77
14/04/04	22/04/04	ADZ	1.518	1.543	6588	164.04
30/04/04	10/05/04	ALZ	1.654	1.618	6048	-217.73
5/05/04	13/05/04	ASX	15.804	15.627	633	-112.23
6/05/04	14/05/04	ARQ	0.857	0.849	11675	-92.23
19/04/04	17/05/04	AMP	5.510	5.560	1815	90.39
11/05/04	19/05/04	ALL	3.913	3.853	2556	-152.59
12/05/04	20/05/04	ANN	7.445	7.642	1343	263.63
20/05/04	28/05/04	AMP	5.616	5.626	1781	17.81
27/05/04	4/06/04	AWB	4.896	4.885	2042	-24.10

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
1/06/04	17/06/04	BIL	5.897	6.050	1696	258.64
15/06/04	23/06/04	AIZ	0.370	0.368	27049	-43.28
16/06/04	24/06/04	ASX	15.581	15.351	642	-147.21
21/06/04	1/07/04	APE	7.100	7.202	1408	142.91
22/06/04	30/06/04	ALH	2.392	2.464	4180	297.20
29/06/04	7/07/04	AIZ	0.371	0.375	26983	107.93
17/06/04	9/07/04	BBB	0.288	0.348	34698	2064.53
1/07/04	13/07/04	ADL	0.188	0.155	53220	-1772.23
7/07/04	16/07/04	ALL	5.002	5.147	1999	290.65
2/07/04	19/07/04	AMP	6.284	6.312	1591	45.34
19/07/04	21/07/04	BCL	0.385	0.318	25988	-1738.60
16/07/04	26/07/04	ANZ	17.941	17.863	557	-43.45
23/07/04	2/08/04	BBL	0.120	0.125	83333	416.67
30/07/04	9/08/04	AIZ	0.372	0.359	26853	-370.57
3/08/04	11/08/04	ASX	15.833	15.196	632	-402.33
4/08/04	12/08/04	AUN	0.929	0.885	10767	-469.44
9/08/04	17/08/04	AXA	3.866	3.902	2587	93.13
9/08/04	17/08/04	ANZ	17.890	17.813	559	-43.04
13/08/04	23/08/04	ALL	5.463	5.469	1831	11.54
23/08/04	31/08/04	AMP	6.189	6.193	1616	7.11
2/09/04	7/09/04	BHP	13.287	13.354	753	50.83
31/08/04	8/09/04	ALN	7.306	7.581	1369	376.20
3/09/04	13/09/04	AIZ	1.760	1.784	5682	138.64
1/09/04	14/09/04	BIL	7.030	7.186	1422	221.97
24/08/04	16/09/04	ASB	1.372	1.437	7287	472.93
15/09/04	23/09/04	BLD	6.900	6.764	1449	-198.22
9/09/04	24/09/04	ANZ	18.618	18.662	537	23.79
17/09/04	27/09/04	BKL	12.025	12.612	832	488.63
20/09/04	28/09/04	BHP	13.931	13.867	718	-46.53
20/09/04	28/09/04	ANC	9.823	9.500	1018	-328.71
21/09/04	29/09/04	ALN	7.290	7.511	1372	302.53
23/09/04	30/09/04	AVX	0.274	0.230	36550	-1604.55
27/09/04	5/10/04	BBB	0.372	0.358	26867	-384.20
5/10/04	6/10/04	ALH	3.229	3.232	3097	9.91
30/09/04	8/10/04	BHP	14.430	14.866	693	301.52
7/10/04	15/10/04	ADB	9.059	9.189	1104	142.86
8/10/04	18/10/04	ASX	16.465	17.031	607	343.32
12/10/04	20/10/04	AGL	13.390	13.267	747	-91.43
11/10/04	28/10/04	BBB	0.347	0.350	28818	86.45
25/10/04	2/11/04	BBG	10.670	10.481	937	-176.53
26/10/04	3/11/04	BHP	13.897	14.005	720	77.26

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
28/10/04	5/11/04	ALL	8.705	8.886	1149	208.20
2/11/04	10/11/04	ANZ	20.499	19.960	488	-263.08
11/11/04	19/11/04	AMP	7.140	6.818	1401	-452.10
12/11/04	22/11/04	AEO	1.623	1.597	6163	-156.54
23/11/04	24/11/04	BKV	0.321	0.229	31192	-2869.66
17/11/04	25/11/04	BIL	7.102	7.053	1408	-70.12
18/11/04	29/11/04	BBB	0.366	0.445	27330	2167.27
18/11/04	29/11/04	BBB	0.366	0.445	27330	2167.27
25/11/04	6/12/04	ANC	10.200	9.757	980	-434.04
1/12/04	8/12/04	APV	0.530	0.430	18868	-1894.35
6/12/04	14/12/04	BHP	15.256	15.122	655	-88.10
6/12/04	14/12/04	ANZ	20.169	20.004	496	-81.79
13/12/04	21/12/04	AMP	6.832	7.116	1464	414.90
8/12/04	29/12/04	BNB	9.980	10.484	1002	504.81
10/12/04	30/12/04	AXM	0.120	0.125	83333	416.67
15/12/04	12/01/05	BBB	0.469	0.476	21317	153.48
4/01/05	12/01/05	ASX	20.669	21.027	484	173.32
12/01/05	17/01/05	AUO	0.623	0.749	16041	2013.15
11/01/05	19/01/05	AMP	7.386	7.736	1354	473.90
17/01/05	25/01/05	ABS	5.593	5.545	1788	-85.11
27/01/05	4/02/05	ASX	21.248	21.060	471	-88.31
7/02/05	15/02/05	AMP	7.850	7.558	1274	-372.65
8/02/05	16/02/05	ARG	5.585	5.725	1791	250.92
10/02/05	18/02/05	ANN	9.745	9.971	1026	231.67
21/02/05	1/03/05	BNB	10.753	10.678	930	-69.56
22/02/05	2/03/05	ALL	10.645	10.634	939	-11.08
4/02/05	4/03/05	BBB	0.608	0.549	16453	-975.66
25/02/05	7/03/05	AGL	13.863	14.062	721	143.70
8/03/05	16/03/05	BNB	11.013	11.035	908	19.34
21/03/05	29/03/05	AZR	0.295	0.271	33910	-796.89
23/03/05	31/03/05	APZ	0.206	0.210	48544	194.18
31/03/05	8/04/05	AMP	7.113	6.875	1406	-334.77
4/04/05	12/04/05	ABS	5.616	5.860	1781	434.74
8/04/05	18/04/05	BGF	0.241	0.209	41545	-1325.29
14/02/05	21/04/05	BHP	16.887	16.832	592	-32.38
19/04/05	28/04/05	BBB	0.422	0.431	23725	220.64
29/04/05	9/05/05	ANZ	21.623	21.713	462	41.49
20/04/05	10/05/05	BEN	9.330	9.199	1072	-140.65
3/05/05	11/05/05	BOQ	11.052	10.768	905	-257.38
6/05/05	16/05/05	ABS	5.253	5.010	1904	-463.05
11/05/05	19/05/05	BGF	0.207	0.208	48239	14.47

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
17/05/05	31/05/05	ASX	19.858	22.933	504	1549.85
26/05/05	3/06/05	ABS	5.330	5.024	1876	-573.49
31/05/05	8/06/05	AMP	6.631	6.623	1508	-11.61
1/06/05	9/06/05	BHP	16.809	17.137	595	195.52
24/05/05	10/06/05	BBB	0.381	0.385	26254	115.52
27/05/05	14/06/05	BNB	11.604	13.354	862	1508.76
15/06/05	20/06/05	ARX	0.046	0.064	215983	3714.91
14/06/05	22/06/05	AEO	1.730	1.711	5781	-108.10
15/06/05	23/06/05	BKN	2.719	2.836	3677	429.11
20/06/05	28/06/05	ABS	5.204	5.445	1921	462.77
22/06/05	30/06/05	BIL	8.303	8.131	1204	-207.33
1/07/05	12/07/05	BBB	0.380	0.353	26316	-721.06
8/07/05	18/07/05	BEN	9.696	9.781	1031	87.43
5/07/05	19/07/05	AGL	14.544	14.392	688	-103.96
14/07/05	20/07/05	BHP	18.743	18.487	534	-136.44
12/07/05	20/07/05	AOE	0.456	0.479	21925	491.12
19/07/05	2/08/05	ASX	23.564	25.781	424	939.97
25/07/05	2/08/05	AXA	4.464	4.479	2240	33.60
29/07/05	9/08/05	BHP	19.379	20.264	516	456.56
1/08/05	9/08/05	AMP	6.770	6.762	1477	-11.08
5/08/05	12/08/05	AXI	0.020	0.024	500000	2000.00
5/08/05	15/08/05	ALN	9.756	9.973	1025	222.32
9/08/05	17/08/05	BBB	0.373	0.349	26788	-640.23
4/07/05	19/08/05	BNB	13.939	16.266	717	1668.03
11/08/05	19/08/05	BHP	20.619	20.575	485	-21.44
22/08/05	24/08/05	BCA	3.786	3.785	2642	-0.79
23/08/05	31/08/05	AXA	4.816	4.891	2076	155.70
31/08/05	1/09/05	ANZ	22.011	22.253	454	110.19
29/08/05	6/09/05	BHP	20.184	20.218	495	16.73
8/09/05	16/09/05	ANZ	22.808	23.165	438	156.54
6/09/05	20/09/05	BNB	18.553	18.484	539	-37.14
18/08/05	29/09/05	ALN	10.705	11.355	934	606.91
15/09/05	3/10/05	BCA	3.608	3.467	2772	-390.02
28/09/05	6/10/05	ANZ	24.083	23.393	415	-286.31
29/09/05	7/10/05	AWB	5.154	5.125	1940	-54.71
30/09/05	10/10/05	AXA	4.870	4.636	2053	-481.22
7/10/05	20/10/05	BDG	1.081	1.078	9254	-26.84
6/10/05	20/10/05	BNB	16.646	16.121	601	-315.71
24/10/05	26/10/05	ANZ	23.565	23.344	424	-93.96
21/10/05	31/10/05	ALL	10.898	11.509	918	560.16
24/10/05	1/11/05	BLD	7.647	7.604	1308	-56.24

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
27/10/05	4/11/05	BBB	0.340	0.340	29438	17.66
4/11/05	7/11/05	AWB	5.282	5.277	1893	-9.28
8/11/05	16/11/05	ANZ	23.141	23.810	432	289.09
9/11/05	17/11/05	AGL	15.049	15.501	665	300.58
16/11/05	24/11/05	ALL	11.875	12.472	842	502.42
17/11/05	25/11/05	BNB	18.037	17.621	554	-230.57
22/11/05	30/11/05	ANZ	23.756	24.015	421	108.74
24/11/05	2/12/05	BHP	21.807	21.926	459	54.62
9/12/05	12/12/05	BHP	21.652	21.888	462	108.75
5/12/05	13/12/05	BIL	9.963	9.935	1004	-27.81
6/12/05	14/12/05	ALN	10.877	10.864	919	-11.86
12/12/05	20/12/05	ANZ	23.191	23.827	431	274.16
13/12/05	21/12/05	BNB	16.673	17.028	600	212.88
14/12/05	22/12/05	ASX	30.254	31.532	331	422.89
19/12/05	3/01/06	APZ	1.197	1.242	8357	379.41
20/12/05	3/01/06	ALN	11.194	11.089	893	-94.57
3/01/06	11/01/06	BBG	14.428	14.062	693	-253.64
13/01/06	17/01/06	BLN	0.248	0.286	40258	1505.65
11/01/06	19/01/06	BHP	23.867	24.191	419	135.50
16/01/06	23/01/06	AXI	0.066	0.086	151515	3015.15
23/01/06	31/01/06	ABS	7.488	7.409	1336	-105.14
10/01/06	10/02/06	BNB	17.538	16.820	570	-408.92
16/02/06	24/02/06	BHP	24.115	24.598	415	200.57
20/02/06	3/03/06	AEO	1.611	1.690	6207	488.49
24/02/06	13/03/06	AGL	19.162	18.743	522	-218.40
7/03/06	15/03/06	ASL	1.634	1.585	6120	-300.49
28/03/06	29/03/06	ASX	34.359	33.315	291	-304.04
16/03/06	31/03/06	BNB	17.741	18.390	564	365.81
27/03/06	4/04/06	BIL	10.554	10.904	947	330.69
20/03/06	7/04/06	AOE	0.626	0.634	15987	127.90
30/03/06	7/04/06	AGL	18.567	18.629	539	33.31
3/04/06	12/04/06	BHP	28.711	29.681	348	337.56
14/03/06	18/04/06	AUN	1.142	1.177	8760	307.48
1/05/06	4/05/06	AWB	4.580	4.400	2183	-394.25
11/05/06	19/05/06	BNB	20.308	18.928	492	-678.57
24/05/06	1/06/06	BBB	0.101	0.087	99305	-1370.41
25/05/06	2/06/06	AWB	4.238	4.466	2360	539.26
26/05/06	5/06/06	ALZ	2.000	2.056	4999	279.94
9/05/06	6/06/06	ACR	0.799	0.777	12517	-271.62
5/06/06	13/06/06	BBB	0.085	0.081	118203	-437.35
6/06/06	15/06/06	BNB	20.585	19.658	486	-450.33

Table B.1 – continued from previous page

Open Date	Close Date	Ticker	Opening ($)	Closing ($)	Units	Profit/Loss ($)
8/06/06	16/06/06	ABS	6.863	6.677	1457	-271.15
9/06/06	19/06/06	BHP	26.472	27.023	378	208.43
13/06/06	21/06/06	AMP	8.489	8.763	1178	323.48
19/06/06	27/06/06	BIL	10.729	10.573	932	-145.02
28/06/06	6/07/06	ANZ	25.848	26.613	387	295.90
30/06/06	10/07/06	BHP	28.878	28.879	346	0.42
3/07/06	11/07/06	BIL	10.832	10.825	923	-6.46
4/07/06	12/07/06	ADZ	2.569	2.578	3893	33.87
7/07/06	17/07/06	ANZ	26.848	24.821	372	-754.19
23/06/06	20/07/06	BNB	20.216	20.030	495	-91.92
13/07/06	21/07/06	AGL	18.180	19.064	550	486.42
19/07/06	27/07/06	APZ	1.418	1.466	7052	340.61
5/07/06	2/08/06	ASX	33.117	32.029	302	-328.70
25/07/06	2/08/06	AMP	9.053	8.790	1105	-290.73
26/07/06	8/08/06	BHP	27.915	27.638	358	-99.35
9/08/06	11/08/06	AWB	3.525	3.414	2837	-314.34
4/08/06	15/08/06	AVA	0.065	0.062	152905	-519.88
8/08/06	16/08/06	ARG	7.140	7.353	1401	298.13
15/08/06	23/08/06	BBB	0.095	0.100	104822	450.73
5/09/06	14/09/06	BAX	3.620	3.636	2762	43.92
6/09/06	14/09/06	ALZ	1.811	1.846	5523	193.31
18/09/06	26/09/06	APE	7.900	8.000	1266	126.60
26/09/06	6/10/06	BBB	0.155	0.155	64475	-6.45
17/10/06	26/10/06	BGF	0.274	0.250	36456	-878.59

Table B.1: A Full Trading Summary print-out, showing the open & close information for the simulation run with a 7 day trading window, and no short-selling. Over the period 1995-2006, the simulator makes 759 trades, with an average annualised return of 0.76%.

Ticker	Trades	Return	Ticker	Trades	Return	Ticker	Trades	Return
ARX	1	37.15%	ALH	2	1.54%	BOQ	11	-1.16%
BGN	1	33.63%	BNB	14	1.43%	ADZ	7	-1.43%
BNO	1	26.11%	ARG	7	1.28%	BCA	5	-1.61%
AXI	8	21.12%	ASX	36	1.24%	ASL	3	-1.68%
BLN	1	15.06%	ALN	12	0.99%	BCD	3	-1.79%
AUO	3	14.80%	BIL	25	0.72%	BKW	2	-2.04%
ADA	3	7.94%	APE	5	0.66%	AGI	5	-2.22%
AGM	2	7.36%	AXA	19	0.57%	ABC	1	-2.35%
BDS	3	4.55%	ADB	11	0.56%	AEX	1	-2.44%
BKN	1	4.29%	BHP	139	0.46%	AHO	1	-2.63%
AXM	1	4.17%	BAX	1	0.44%	ACR	1	-2.72%
BBL	1	4.17%	NAB	21	0.33%	AUN	18	-3.14%
ALL	18	3.56%	ANZ	73	0.32%	AVA	1	-5.20%
AVJ	6	3.11%	AWB	13	0.27%	BGF	7	-5.39%
AOE	2	3.10%	AIZ	13	0.27%	BDL	2	-6.74%
APZ	3	3.05%	ALZ	25	0.24%	AZR	1	-7.97%
BLD	8	2.91%	AMP	54	0.19%	AHD	1	-9.53%
AQA	1	2.79%	AGL	25	0.19%	BLT	1	-12.85%
BBB	29	2.69%	ASB	4	0.06%	AVX	1	-16.05%
ADG	2	2.41%	ANC	4	-0.18%	AET	2	-17.03%
ANN	3	2.37%	BDG	1	-0.27%	BCL	1	-17.39%
BKL	3	2.31%	ANE	3	-0.27%	ADL	1	-17.72%
ABS	10	2.20%	BKR	2	-0.32%	APV	1	-18.94%
AEO	5	1.82%	ANM	1	-0.49%	ASC	1	-21.81%
BPC	10	1.79%	BBG	9	-0.56%	BKV	1	-28.70%
BEN	27	1.57%	ARQ	4	-0.72%			

TABLE B.2: Trading Returns by Company, using a 7 day trading window and no short-selling, with an average annualised return of 0.76%.

Ticker	Trades	Return	Ticker	Trades	Return	Ticker	Trades	Return
AIZ	10	46.68%	ASX	23	7.21%	ABC	1	0.24%
ARX	1	37.15%	BBL	2	6.88%	ANN	3	-0.65%
BGN	1	33.63%	BBG	7	6.73%	ASB	4	-0.76%
ADA	3	26.36%	BGF	6	6.61%	AEO	4	-3.13%
BNO	1	26.11%	AOE	2	6.45%	ADZ	6	-3.47%
AUO	3	19.95%	ANC	2	5.87%	ACR	1	-3.71%
BDG	1	19.84%	BLD	8	5.82%	BCA	5	-4.13%
BKN	1	19.66%	AGL	13	5.58%	AVA	1	-5.20%
AXM	1	19.33%	BPC	9	4.44%	ADG	2	-6.24%
AEX	1	18.90%	ANZ	40	4.00%	AGI	4	-6.81%
BAX	1	17.68%	ADB	9	3.51%	BCD	3	-7.37%
AXI	7	16.68%	BIL	17	3.49%	AHO	1	-7.89%
BLN	1	15.06%	AQA	1	3.12%	BKR	2	-8.71%
BNB	7	13.35%	ARQ	4	2.99%	BKW	2	-10.03%
BKL	3	12.60%	BDS	2	2.24%	AVX	1	-16.05%
APZ	3	11.90%	AVJ	6	2.19%	AHD	1	-16.33%
NAB	7	11.52%	BDL	2	2.13%	BLT	1	-16.37%
ABS	5	9.97%	ARG	6	1.74%	BCL	1	-17.39%
ALL	16	9.12%	ANE	3	1.72%	ADL	1	-17.72%
ALH	2	8.64%	AXA	13	1.70%	APV	1	-18.94%
ALN	6	8.30%	AUN	14	1.62%	AZR	1	-18.96%
BEN	19	7.85%	AWB	11	1.42%	ANM	1	-19.04%
APE	4	7.80%	BHP	91	1.05%	ASC	1	-21.81%
ASL	3	7.63%	AMP	28	0.94%	AET	2	-23.69%
ALZ	12	7.50%	BOQ	5	0.50%	BKV	1	-28.70%
AGM	2	7.36%	BBB	23	0.47%	AHC	1	-40.00%

TABLE B.3: Trading Returns by Company, using a 90 day trading window and no short-selling with an average annualised return of 4.27%.

B2 Long/Short Strategy

Ticker	Trades	Return	Ticker	Trades	Return	Ticker	Trades	Return
ARX	1	37.15%	AEO	11	0.99%	BCA	11	-0.90%
BGN	1	33.63%	AMP	75	0.96%	BOQ	13	-0.94%
BNO	1	26.11%	AIZ	19	0.70%	AUN	24	-1.12%
AZA	1	18.55%	APE	5	0.66%	ADZ	15	-1.26%
BDM	1	18.07%	ANN	4	0.63%	BCD	5	-1.42%
AXI	9	16.83%	ADB	11	0.56%	BBG	12	-1.46%
BLN	1	15.06%	AVJ	12	0.46%	BPC	24	-1.48%
AUO	4	13.78%	BAX	1	0.44%	ASL	3	-1.68%
AVR	2	8.07%	BHP	199	0.43%	BKW	2	-2.04%
BOC	2	7.98%	ARP	1	0.42%	AVM	1	-2.28%
ADA	3	7.94%	NAB	23	0.40%	AEX	1	-2.44%
AGM	2	7.36%	BDS	6	0.36%	AHO	1	-2.63%
ABC	4	5.21%	BKV	4	0.35%	ACR	1	-2.72%
BKN	1	4.29%	AXA	30	0.31%	BGF	10	-3.78%
AXM	1	4.17%	ALZ	26	0.31%	AGI	7	-4.03%
AOE	2	3.10%	ABB	1	0.27%	AVA	1	-5.20%
APZ	3	3.05%	BEN	36	0.22%	BDL	2	-6.74%
BBB	36	2.64%	ARQ	7	0.15%	BBL	4	-7.84%
ADG	2	2.41%	BIL	43	0.13%	AZR	1	-7.97%
BKL	3	2.31%	ANZ	110	0.09%	AHD	1	-9.53%
BNB	15	1.90%	ASB	4	0.06%	BLT	1	-12.85%
BLD	15	1.66%	AGL	32	-0.06%	AVX	1	-16.05%
AWB	28	1.55%	ANC	4	-0.18%	AET	2	-17.03%
ALL	32	1.51%	BDG	1	-0.27%	BCL	1	-17.39%
APA	1	1.32%	BKR	2	-0.32%	ADL	1	-17.72%
ALH	5	1.29%	ANE	11	-0.46%	APV	1	-18.94%
ASX	54	1.24%	ABS	17	-0.49%	ASC	1	-21.81%
ALN	16	1.13%	ANM	1	-0.49%	BAS	1	-25.50%
ARG	8	1.03%	AQA	2	-0.74%			

TABLE B.4: Trading Returns by Company, using a 7 day trading window and with short-selling with an average annualised return of 0.57%.

Ticker	Trades	Return	Ticker	Trades	Return	Ticker	Trades	Return
AIZ	14	38.59%	ANC	2	5.87%	ABC	4	-2.34%
ARX	1	37.15%	ADB	9	3.51%	ACR	1	-3.71%
BGN	1	33.63%	AWB	19	3.49%	BCA	10	-4.30%
ADA	3	26.36%	AVJ	10	3.26%	AGI	6	-4.62%
BNO	1	26.11%	ASX	33	3.25%	BCD	5	-4.78%
BDG	1	19.84%	BPC	19	3.15%	BBL	4	-4.89%
BKN	1	19.66%	AQA	1	3.12%	ANN	4	-5.01%
AXM	1	19.33%	BIL	25	2.94%	AVA	1	-5.20%
AEX	1	18.90%	ANZ	53	2.84%	AEO	7	-5.24%
AZA	1	18.55%	BGF	8	2.61%	ADG	2	-6.24%
BDM	1	18.07%	ARQ	6	2.27%	AHO	1	-7.89%
BAX	1	17.68%	ANE	10	2.26%	ABS	10	-7.95%
BOC	2	15.33%	BDL	2	2.13%	BKR	2	-8.71%
ARP	1	15.14%	AGL	15	2.05%	BKW	2	-10.03%
BLN	1	15.06%	AMP	39	1.98%	AVX	1	-16.05%
BNB	7	12.91%	BHP	123	1.40%	AHD	1	-16.33%
BKL	3	12.60%	AUN	20	1.35%	BLT	1	-16.37%
AXI	8	12.41%	BLD	12	1.02%	BCL	1	-17.39%
APZ	3	11.90%	ALH	4	0.89%	AVM	1	-17.69%
NAB	7	11.52%	ARG	7	0.46%	ADL	1	-17.72%
AUO	4	10.80%	ALN	10	0.41%	APV	1	-18.94%
ALL	26	8.37%	AVR	2	0.20%	AZR	1	-18.96%
BEN	22	8.23%	BDS	4	-0.01%	ABB	1	-19.02%
ALZ	13	8.17%	BBG	10	-0.26%	ANM	1	-19.04%
APE	4	7.80%	ADZ	10	-0.69%	ASC	1	-21.81%
ASL	3	7.63%	ASB	4	-0.76%	AET	2	-23.69%
AGM	2	7.36%	BBB	27	-0.89%	BAS	1	-25.50%
AOE	2	6.45%	AXA	18	-1.33%	AHC	1	-40.00%
BKV	4	6.03%	BOQ	6	-2.28%			

TABLE B.5: Trading Returns by Company, using a 90 day trading window and with short-selling and an average annualised return of 3.15%.

Bibliography

[1] A. Devitt and K. Ahmad, "Sentiment polarity identification in financial news: A cohesion-based approach," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 984–991, Association for Computational Linguistics, June 2007.

[2] J. M. Wiebe, "Tracking point of view in narrative," *Computational Linguistics*, vol. 20, no. 2, pp. 233–287, 1994.

[3] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *WWW '03: Proceedings of the 12th international conference on World Wide Web*, (New York, NY, USA), pp. 519–528, ACM, 2003.

[4] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *In Proceedings of IEEE International Conference on Data Mining (ICDM*, pp. 427–434, 2003.

[5] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which side are you on?: identifying perspectives at the document and sentence levels," in *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*, (Morristown, NJ, USA), pp. 109–116, Association for Computational Linguistics, 2006.

[6] R. Schumaker, Y. Zhang, and C. Huang, "Sentiment analysis of financial news articles," in *Proceedings of the 20th Annual Conference of International Information Management Association*, (Houston, TX), 2009.

[7] S. R. Das, F. d. A. Martinez-Jerez, and P. Tufano, "eInformation: A Clinical Study of Investor Discussion and Sentiment," *Financial Management, Vol. 34, No. 3, pp. 103-137, Autumn 2005*, 2005.

[8] A. Mowshowitz, "On the market value of information commodities. i. the nature of information and information commodities," *Journal of the American Society for Information Science*, vol. 43, no. 3, pp. 225–232, 1992.

[9] S. Clarke, "Reuters announces news sentiment analysis to give trading machines market insight," April 2007.

[10] Z. Jelveh, "How a computer knows what many managers don't."

[11] R. Schumaker and H. Chen, "Evaluating a news-aware quantitative trader: The effects of momentum and contrarian stock selection strategies," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 2, pp. 247–255, 2008.

[12] E. Fama, "Random walks in stock market prices," *Financial Analysts Journal*, vol. 21, no. 5, pp. 55–59, 1965.

[13] B. G. Malkiel, "The efficient market hypothesis and its critics," *The Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59–82, 2003.

[14] R. Brealey, S. Myers, and F. Allen, *Principles of Corporate Finance*. McGraw Hill, 9th edition ed., 2008.

[15] R. Ball and P. Brown, "An empirical evaluation of accounting income numbers," *Journal of Accounting Research*, vol. 6, no. 2, pp. 159–178, 1968.

[16] D. Kahneman, "Maps of bounded rationality: Psychology for behavioral economics," *The American Economic Review*, vol. 93, no. 5, pp. 1449–1475, 2003.

[17] J. Lakonishok, I. Lee, and A. M. Poteshman, "Investor Behavior in the Option Market," *SSRN eLibrary*, 2004.

[18] L. Gillam, "Sentiment analysis and financial grids," in *Workshop on Bridging quantitative and qualitative methods for social sciences using text mining techniques*, 2006.

[19] L. J. Brinton, *The Structure of Modern English*. John Benjamins Publishing Company, 2000.

[20] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, (New York, NY, USA), pp. 70–77, ACM, 2003.

[21] N. Godbole, S. M., and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[22] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, (Morristown, NJ, USA), pp. 79–86, Association for Computational Linguistics, 2002.

[23] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. A. Keim, "Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008," in *Visual Interfaces to the Social and the Semantic Web (VISSW 2009)*, Universität Konstanz, 2009.

[24] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 341–349,

ACM, 2002.

[25] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 168–177, ACM, 2004.

[26] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 417–424, Association for Computational Linguistics, 2002.

[27] L. W. Ku, Y. T. Liang, and H. H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[28] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, (New York, NY, USA), pp. 617–624, ACM, 2005.

[29] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, 2008.

[30] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pp. 417–422, 2006.

[31] F. Media, "The australian financial review: Brand overview."

[32] C. Chua, M. Milosavljevic, and J. Curran, "A sentiment detection engine for internet stock message boards," in *The Australasian Language Technology Workshop (ALTW)*, 2009.

[33] J. R. Curran, S. Clark, and J. Bos, "Linguistically motivated large-scale nlp with c&c and boxer," in *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, (Morristown, NJ, USA), pp. 33–36, Association for Computational Linguistics, 2007.

[34] R. M. W. Dixon, *Where have all the adjectives gone? and other essays in semantics and syntax*. Mouton, 1982.

[35] I. Marshall, *The Computational analysis of English: a corpus-based approach*. No. 42-65, Longman, London., 1987.

[36] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[37] E. Charniak, "Statistical techniques for natural language parsing," *AI Magazine*, 1997.

[38] K. Ajdukiewicz, *Die syntaktische konnexita ÌĹt*, vol. In Storrs McCall, ed., Polish Logic 1920-1939, 207–231. Oxford: Oxford University Press. Translated from Studia Philosophica, 1, 1-27., 1935.

[39] Y. Bar-Hillel, "A quasi-arithmetical notation for syntactic description," *Language*, vol. 29, pp. 47–58, 1953.

[40] M. Steedman, "Categorial grammar," 1998.

[41] W. J. Hutchins, "Yehoshua bar-hillel: a philosopher's contribution to machine translation," in *Early years in machine translation*, pp. 299–31, John Benjamins, 2000.

[42] E. Charniak, "A maximum-entropy-inspired parser," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, (San Francisco, CA, USA), pp. 132–139, Morgan Kaufmann Publishers Inc., 2000.

[43] D. Chiang and D. M. Bikel, "Recovering latent information in treebanks," in *Proceedings of the 19th international conference on Computational linguistics*, (Morristown, NJ, USA), pp. 1–7, Association for Computational Linguistics, 2002.

[44] S. Clark, J. Hockenmaier, and M. Steedman, "Building deep dependency structures with a wide-coverage ccg parser," in *In Proceedings of the 40th Meeting of the ACL*, pp. 327–334, 2002.

[45] J. Bos and M. Nissim, "An empirical approach to the interpretation of superlatives," in *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, (Morristown, NJ, USA), pp. 9–17, Association for Computational Linguistics, 2006.

[46] J. P. Ng, "Processing sentiments and opinions in text." National University of Singapore, 2010.

[47] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[48] G. Salton, ed., *Automatic text processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1988.

[49] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[50] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[51] VentureLine, "Ventureline's dictionary of accounting terms."

[52] C. Harvey, ed., *The New York Times Dictionary of Money and Investing: The Essential A-to-Z Guide to the Language of the New Market*. Times Books, 2002.

[53] T. Joachims, "Text categorization with suport vector machines: Learning with many relevant features," in *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, (London, UK), pp. 137–142, Springer-Verlag, 1998.

[54] P. Urbach and C. Howson, *Scientific Reasoning: The Bayesian Approach*. Open Court, 1993.

[55] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[56] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.

[57] K. Nigam, J. Lafferty, and A. Mccallum, "Using maximum entropy for text classification," in *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61–67, 1999.

[58] I. Csiszar, "Maxent, mathematics, and information theory," *Maximum Entropy and Bayesian Methods*, vol. 79, pp. 0–35, 1996.

[59] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression." Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August 2004.

[60] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, 2000.

[61] N. Christianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[62] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Maching Learning Research*, vol. 2, pp. 265–292, 2002.

[63] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," in *Machine Learning in Document Analysis and Recognition* (S. Marinai and H. Fujisawa, eds.), vol. 90 of *Studies in Computational Intelligence*, pp. 361–386, Springer Berlin / Heidelberg, 2008.

[64] M. Aitken and A. Frino, "Execution costs associated with institutional trades on the australian stock exchange," *Pacific-Basin Finance Journal*, vol. 4, no. 1, pp. 45 – 58, 1996.

[65] S. M. Kakade, M. Kearns, Y. Mansour, and L. E. Ortiz, "Competitive algorithms for vwap and limit order trading," in *EC '04: Proceedings of the 5th ACM conference on Electronic commerce*, (New York, NY, USA), pp. 189–198, ACM, 2004.

[66] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval* (D. E. Losada and

J. M. Fernández-Luna, eds.), vol. 3408 of *Lecture Notes in Computer Science*, pp. 345–359, Springer Berlin / Heidelberg, 2005.

[67] C. Van Rijsbergen, ed., *Information Retrieval*. 1979.

[68] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, (San Francisco, CA, USA), pp. 1137–1143, Morgan Kaufmann Publishers Inc., 1995.