Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Generating Sentiment Lexica: Evaluating Approaches with Genetic Algorithms and Particle Swarm Optimization

Nicholas Daly

Thesis submitted for the Degree of Doctor of Philosophy

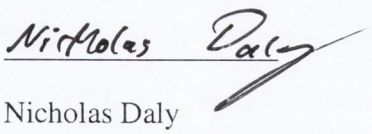
Department of Computer Science and Statistics
University of Dublin
Trinity College

July 2014

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the Universitys open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.



Nicholas Daly

Summary

This thesis examines the application of sentiment analysis towards financial news. The primary goal within the field of sentiment analysis is to develop a methodology by which the sentiment or emotional view being expressed by the author(s) of a text may be extracted and quantified.

Within the financial domain sentiment analysis aims to quantify the polarity of textual data related to financial institutions. Past works have included the examination of news stories, financial fillings, search trends, and message board posts. This data can either be viewed as having an impact on the future financial performance of a company or index whereby the sentiment will either have an impact on investor confidence, or may indicate information which has yet to be incorporated into current market price. An alternative view is that any sentiment extracted will merely be a proxy of currently existing investor sentiment, and while not having any impact on market prices, does provide valuable information regarding views of the market.

I apply a bag of words approach where texts are summarised as a vector of terms and their frequency. Through the examination of this vector the emotional polarity of the text may be extracted. A key aspect of this approach is the identification of appropriate terms to be included for later examination. A common approach includes the usage of comprehensive reference dictionaries, whereby terms are categories according to their polarity, topic and usage.

However, such dictionaries may be flawed in that the specific meaning of a term is domain specific, those from one domain do not translate to all others with the same affect bearing. In this I aim to propose a methodology whereby given an external peg I refine any given list of term in an objective and systematic manner. This is achieved through the usage of two biologically inspired algorithms, *genetic algorithms* and *particle swarm optimization*; I train these against six months of financial news against future market movements. Both of these search heuristics have been shown to be successful within numerous optimization and search problems. I collected a news corpus from the AP financial news wire concerning the US economy, and examine the sentiment extracted using four different lexica against three US market indices. By refining or *pruning* these reference lexica, where terms are selected for inclusion and exclusion, the predictive ability of the affect based time series on future market returns is improved, successfully eliminating terms which within the domain of finance are misclassified in regards to their affect meaning.

A secondary area of examination is the role of news flow in sentiment analysis. The volume of news arrival is non-static and highly volatile with changes in the frequency of news arrivals over time. The larger the discussion taking place may indicate the greater the importance of such an event. Within this I examine the role and frequency of duplicate news items. Where news stories may be repeated either in their entirety or with minor modifications or updates. While these news item may contain no new information the mere act of reprinting and updating a commentary may indicate the importance of an event or perhaps an attempt to drive a discussion towards a given topic or event. I find that the inclusion of such duplicate news items has a significant impact on sentiment analysis and their inclusion significantly improves any statistical analysis against future market movements.

Acknowledgements

Misura ciò che è misurabile, e rendi misurabile ciò che non lo è
-Measure what is measurable, and make measurable what is not so.
-Galileo Galilei

There are a number of people without whos assistance and support this work would not have been possible. Firstly I would like to offer my sincere thanks to professor Khurshid Ahmad for his support and guidance and giving me to opportunity to carry out this work. His wide and extensive knowledge has provided a continuous source for the discussion of topics and areas allowing for a greater exploration and discussion of various areas of sentiment analysis, linguistics, and cognitive reasoning. I would also like to thank the school of computer science and statistics and the Faireacháin project for their generous funding.

This thesis is dedicated to my parents, without whom none of this would be possible. During my life they have given me nothing less than their complete and unwavering support. The examples they have set, their encouragement and belief in me has made all this possible. To Paul, Dee, Elizabeth, Dominik and Alice I would like to thank you for always cheering me on and rooting for me, and for knowing that I could manage this. To my nephews Conor, John and Aidan thank you simply for giving me a regular chance to forget about things, and a constant excuse to just act foolish.

Lastly to Steph; thank you for your constant support during all these years, for always being there to listen, for knowing when to tell me it will be ok, when to tell me to relax and when to simply nod along as I talked. Most of all thank you for riding out this awkward, at times ridiculous but always fantastic journey. Lastly thank you for saying yes.

Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

The role of sentiment and affect on financial market is an area of growing interest. Increasingly researchers are interested in the role which sentiment has on financial movements. Investors are susceptible to sentiment towards the financial markets, market increases may be described as *bullish* a period of "characterized optimism, investor confidence and expectations"[1] or *bearish* where "prices of securities are falling, and widespread pessimism causes the negative sentiment to be self-sustaining".[2] In these definitions changes are cited as the result of optimism and pessimism, perhaps not rooted in technical information but in the mood of market sentiment. If such movements are influence through sentiment within the market, one should ask where might this be found? One post hoc view is that contemporary reports capture the behaviour and aspirations of the stakeholders. This idea led some to believe that if sentiment is presented within such documents it may be supposed that it is presented in a manner to influence a reader. So as to produce this emotional response the author will attempt to convey such information through the usage of specific terminology; thereby sentiment is *encoded* within texts.

The extraction of such terminology allows for an abstraction of complex emotions; producing frequency measurements of affect bearing terms. In this manner unstructured qualitative data within language may be abstracted to produce structured quantitative information. This presents the challenge of determining which of these terms provide the greatest level of sentiment bearing. This problem may be approached in a number of fashions; a human *intuition* basis, individuals with knowledge of the field produce lists of terms which according to their knowledge and views capture the language of the domain. A linguistic approach; conducting experiment asking individuals to identify the affect of a given term. Or from a *Corpora* examination approach; from examining the structure and terminology within a representative corpus and identifying frequently occurring terms. As *"Assessments by native speakers of the relative acceptability largely correlates with their assessments of the relative frequency" [83],*

[1] http://www.investopedia.com/terms/b/bullmarket.asp (last accessed 31/07/2013)
[2] http://www.investopedia.com/terms/b/bearmarket.asp (last accessed 31/07/2013)

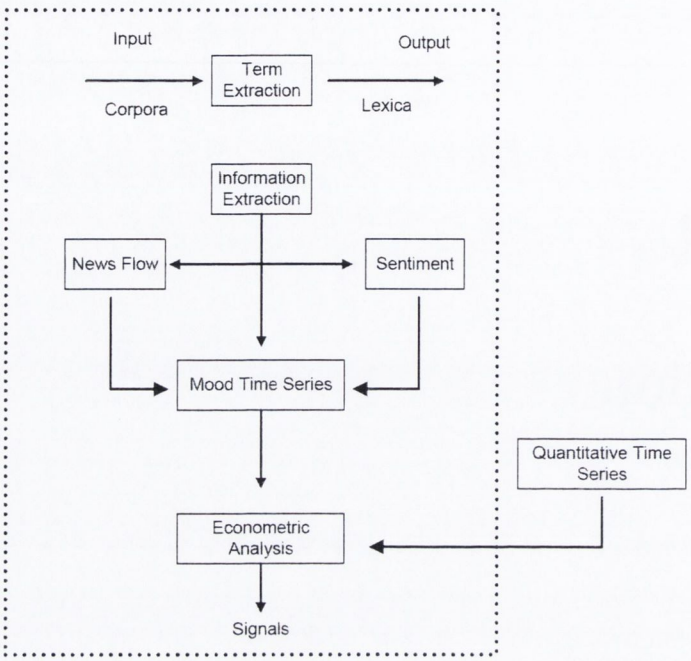


Figure 1.1: Qualitative Information Extraction & Econometric Integration.

the repetition of a term within a domain represents its acceptance. From this we have the method for the potential examination of sentiment within a given text and once quantified may track its movements over time.

1.1 Motivation

The first motivation for this research stems from the increase in the availability of financial news and information and its value towards financial analysis. There appears to be no systematic way in which such data is curated and the information extracted from it. The growth in information content available through the internet; message boards, news sites and web forums has lead to an information overload.

The ability to develop a system which allows for the automated concurrent processing of financial data and news is currently a key challenge of the 21^{st} century. Within this area I see the challenge of identifying and processing potential empirical content of news in conjunction with numerical data as very important. As has been noted when examining affect term frequency the target domain must be taken into consideration, financial news reporting may contain a number of linguistical variations between what is observed in other areas. While in theory financial reporting should be of a factual and analytical form, such an area is rich with regards to metaphorical language. Discussions on the *health* of a company, *mood* of the market, and the risk of *contagion* are common within financial news.

An examination of the contents and flow of news items allows for an evaluation of the frequency of key affect terms as defined against a given lexicon. By counting the frequency of a given list of affect terms in a corpus of news items an affect score for each day can be calculated. The frequency of a given list of affect terms is summed across all documents on a given day and a sentiment score is calculated for the day. By ordering these affect scores in a chronological manner an *affect time series* can be constructed. This time series outlines the changes of the sentiment measurement within a the retrieved news corpus over time. This allows for the development of a structured *mood time series* from such unstructured emotional content. For economic analysis, this qualitative information may be examined against quantitative data such as price movements; the goal of which being to combine the two as shown in figure 1.1.

The second motivation stems from the reliance on *legacy lexica*; researches examining the emotional content contained within texts frequently rely on precompiled lists of positive and negative affect terms used by speakers of English, German or indeed any natural language. While these lexica are valuable tools within textual analysis they may fail to capture the nature of the language of the domain being evaluated. In financial language the term *share* while positive in general language has no bearing or emotional content within financial news. Similarly a term such as *crude* while negative in most usages when used in reference to *crude oil* again has no emotional bearing. Crude oil is a physical term, a state of matter to describe a commodity that is traded on a frequent basis. Researchers examining financial news analysis have discussed a consistent manner in which such lexica are developed and while noting the errors; proceed to develop their own term list based on their conclusions. While domain specific lexica may be considered to being superior, their construction is time consuming, and there is no guarantee that the researchers' selection accurately captures the nature of the domain's language.

The goal of this work it to develop a system from which a *Contemporary Lexicon* may be developed. Allowing for the automated selection of affect terms when applied to a given domain or topic. Through minimal human interaction a given list of terms may be refined producing terms which when applied to financial news have a negative or positive bearing. These *Contemporary Lexica* are applied to the domain of sentiment analysis of finance. To investigate the potential for an automated extraction of sentiment within financial news and examine the impact towards market behaviour.

This is achieved through the usage of biologically inspired algorithms. From which a given lexicon may be refined; according to some performance measurement, removing terms which are misclassified within the domain of financial news. As outlined in figure 1.1 such a methodology allows for the production of an affect based time series which may be examined against a financial instrument.

While this methodology was developed specifically for an examination of sentiment within

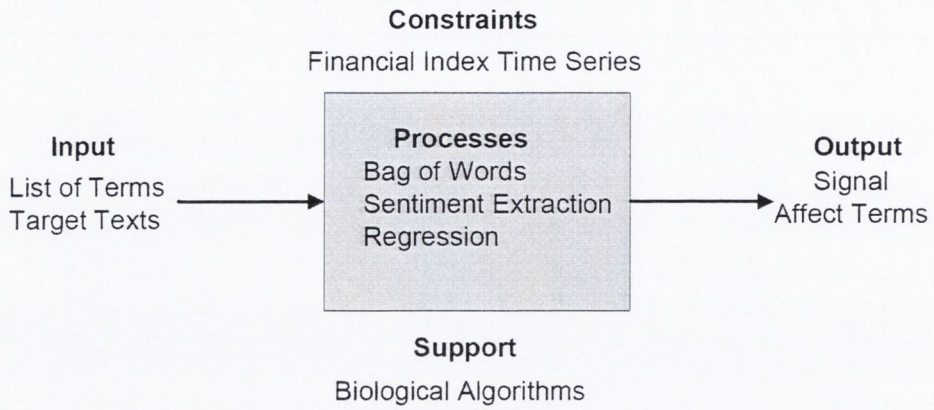


Figure 1.2: Proposed Process for Lexicon Refinement System.

financial markets , with the provision of an adequate performance metric such an approach should transfer to other areas of textual analysis. The nature of the domain, initially provided key terms and texts should be irrelevant; provided there is some means for measuring performance a refined lexicon with appropriate affect terms should be produced. Key terms may be identified allowing not only for superior analysis of texts but also an insight into the nature of the specific structure of the given domain's language and terminology.

1.2 Publications

Daly, N. , Ahmad, K. & Kearney, C. (2009) Correlating Market movements with consumer confidence and sentiments: A longitudinal study Text Mining Services 2009 [18].

My first paper examined a correlation between two well established measurements of change in financial markets return and volatility and consumer confidence about nation economies, and sentiment within financial news. The sentiment extracted was collected from Irish news over a ten year period (1995-2005). This was my first attempt to show a relationship between sentiment and future financial movements. Statistical analysis showed that the measurement of sentiment was found to not be randomly distributed.

Ahmad, K. , Daly, N. & Liston, L. (2012) What is new? News media, General Elections, Sentiment, and named entities Sentiment Analysis where AI meets Psychology (SAAIP) 2012 [3].

This paper expanded my work on textual analysis to the area of political news analysis. Using Rocksteady (a program I aided in the prototype development of) we examined if there

was a relationship between candidate citations and performance in the Irish general election of February 2011. News was collected from a number of Irish news sources, both national and regional, and found a relationship between the frequency of candidate names within articles and the election outcome. A follow-up study was conducted on the Irish presidential election and again showed a relationship between candidate citations and election outcome.

1.3 Contributions

Two areas of contribution have been developed during the progress of this work.

Automated Lexicon Refinement

The first and main contribution aims for the development of a methodology allowing for an objective development of a lexicon for the purpose of sentiment analysis. So as to extract sentiment from a text some form of reference dictionary is required; a list of terms and key concepts, identified as being positive or negative in emotional context or indeed any affect category of named entity class. One approach is to select *seed* dictionaries, where a list of terms has been comprehensibly examined, discussed and through manual examination tagged according to the affect polarity. Such dictionaries may be taken whole, regardless of the domain and topic of interest and trusting the comprehensive and systematic manner in which it has been compiled. Alternatively such dictionaries may be used as a starting point and through manual examination and consideration of the domain in question *prune* such dictionaries, eliminating and adding terms based on the researchers own expert knowledge. This filtering of a given dictionary is implemented to account for changes in language, the variation over time in the importance and usage of individual terms, while also being implemented so as to account for the topic domain in question.

Alternatively it may desired to examine the terminology against some external *peg*, examining documents and terminology for the emergence and presence of terms observed during some external event, quantified external variables such as volume of market trades or financial movements. Within the area of sentiment analysis in finance this may be pegged against some market movements, examining which terms occur surround specific market behaviors such as increases or decreases in share prices, trading volume or volatility. Within this work the two methods are combined with the provision of a seed list of terms and an external market index, and *prune* the given list of terms with the aim of identifying the terms which within the given domain contain some affect bearing. This is achieved through the implementation of two forms of biologically inspired algorithms; genetic algorithms and particle swarm optimization.

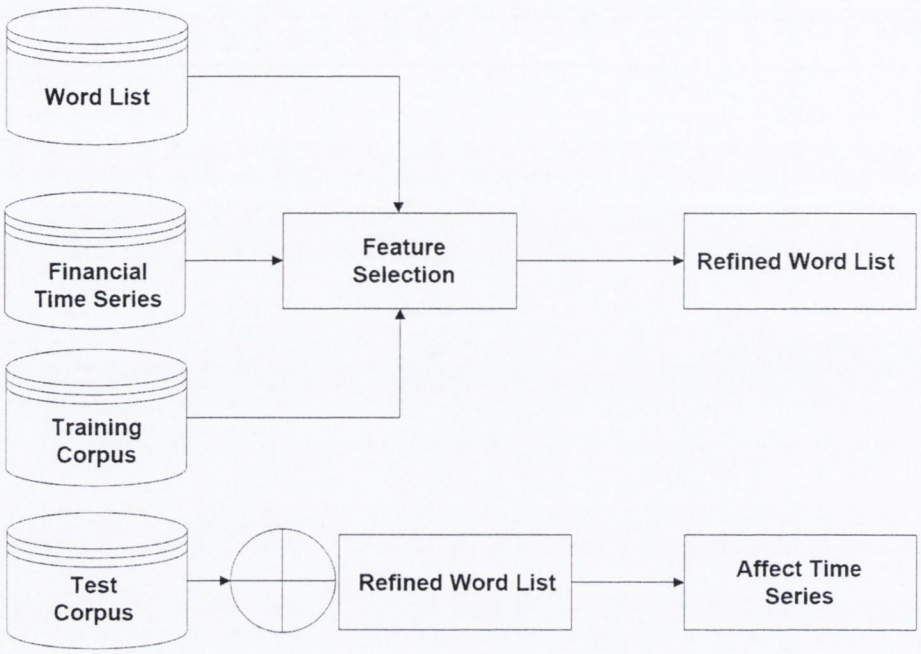


Figure 1.3: Word List Refinement

From this an implementation was developed[3] allowing for the generation of an affect time series, based on the frequency of key affect bearing term according to a given reference lexicon. This resulting time series may be examined against a market index in a statistical manner. The origin of such key terms is irrelevant, the system allows for the refinement of such terms; eliminating those which provide no useful information in regards to the extraction of sentiment specifically applied towards financial language and its bearing on future market returns. In this manner a refined contemporary lexicon is generated which having excluded noise terms and those incorrectly classified as bearing positive or negative affect towards financial news should provide superior explanatory and predictive results on future market returns. In a systematic manner terms may be selected for exclusion and inclusion based on its impact on a developed affect time series as outlined in algorithms 1 and 2, the aim of this is to remove ourselves from selecting and identifying such terms so as to achieve what others have attempted through manual examination, in an automated and objective manner.

This methodology is applied to sentiment analysis of financial news, by collecting news items over a time period and counting the frequency of certain affect terms in news items on each given day a time series recording the frequency of these affect terms over time is constructed. By taking this time series where each data point is the frequency of affect terms and carrying out regression analysis of this time series against a financial index time series I aim to gauge the impact sentiment may have on future market behavior.

[3] All code was developed in Java

When examining movements in financial market it is common to look at the returns of prices, returns can be defined as changes in the natural logarithm of prices. If P_t is the market price of an index or company at time t than the returns at time t R_t can be calculated as

$$R_t = ln(\frac{P_t}{P_{t-1}}) \tag{1.1}$$

When attempting to model returns it is common to examine the influence of past returns and their potential impact on the current market price through regression analysis. If the returns at price t are modeled as a function of the past 5 observed market returns R_t may be modeled as

$$R_t = \alpha + \beta_1(R_{t-1}) + \beta_2(R_{t-2}) + ... + \beta_5(R_{t-5}) \tag{1.2}$$

where α is a constant and β is a coefficient which shows the impact of the value at the given time on R_t. The inclusion of these past value or *lags* can be summarised with an operator $L5$, where $L5(x_t) = [x_{t-1}x_{t-2}x_{t-3}x_{t-4}x_{t-5}]$. This work aims to expand on equation 1.2 and include values from an affect time series into the model so as to investigate if there is a relationship between the affect score collected from a corpus of financial news and future market returns. If $Sent_t$ is the sentiment measurement at time t it is possible to expand the equation to include sentiment as shown in equation 1.3.

$$R_t = \alpha + \beta.L5(R_t) + \gamma.L5(Sent_t) \tag{1.3}$$

While quantitative information is based on market movements, trading volume, and other technical variables, and has been the source of much study and examination regarding their role in the behaviour of the market. The role of such qualitative information is something which in not understood to the same degree.

First, given a list of terms and a training corpus, a time series of term frequency is generated, evaluated, and regressed against a financial time series. Through machine learning techniques terms are selected for inclusion and exclusion, resulting in a new word list of key affect terms. This refined term list is then used for later analysis and against a test corpus for the generation of an affect time series as illustrated in figure 1.3. In regards to the selection of appropriate word lists two varying previously developed lexica are used. Firstly the negative and positive affect categories provided by the General Inquirer developed by Stone et al., this comprehensive and independently developed dictionary provides a valuable resource whereby a number of individuals within a committee in a systematic manner classified thousands of terms across a number of categories. Secondly I examine a dictionary developed specifically for the usage of sentiment analysis of financial news developed by Loughran & McDonald professors of finance at the university of Notre Damn, this dictionary was constructed through manual evaluation of

frequently occurring terms found within financial reports classified according to the authors' knowledge of financial language. These two sources of negative and positive affect word lists allow for an examination of the impact of domain specific dictionaries and provide the advantage that both have been used in a number of studies in the field of sentiment analysis of financial texts and their impact on future market movements. The performance of such a methodology is examined against three previously studies indices within sentiment analysis; the Dow Jones Industrial Average, the S&P 500 and the VIX index. The selection of these indices was that a number of studies from econometricians have concentrated on these as major indices which may reflect the health of the US economy. As such these are examined against financial news regarding the US economy retrieved from the AP financial NewsWire.

News Flow and Data Curation

The second area of research relates to the field of data curation; a system was developed which with minimal human interaction allows for the collection, sanitisation and archiving of corpora. Data collected from an online repository may be ordered and sanitized with extra metadata and markup comment removed. Further text items may be examined against those within the corpus so as to examine if the given news piece is a duplicate of a previously released item. Such duplicates may be verbatim in nature and a simple reprint, or may represent a modification or extension of a previous news piece. A high flow of news arrival, the continuous updating of news commentaries and events may indicate the development of some key event. Updated versions of a news item are released as a story or event evolves and issues, and facts emerge. Such an examination provides three corpora; one containing all the news items, regardless if they contain new information or not, one comprised of no exact duplicates where only those items which have been re-released word for word are excluded and lastly *near duplicates* where news items which are found to be an variation, expansion or correction of a previously released news item are excluded.

1.4 Thesis Roadmap

The remainder of this thesis is presented as follows:

- AN ALGORITHMIC BASE FOR DICTIONARY CONSTRUCTION (CHAPTER 2) Presents a discussion on the field of textual analysis, presenting some of the key resources available and their usage. This is followed by an examination on sentiment analysis and in particular its applications within financial news analysis. Lastly is a discussion on biologically inspired algorithms and present two varying forms; genetic algorithms

and particle swarm optimization. Their methodology and application towards textual analysis form a key topic for discussion.

- METHODS FOR CONTEMPORARY LEXICA DEVELOPMENT (CHAPTER 3) The processes by which news items may be collected, ordered and from which, construct an affect based time series is presented. Through the examination of the frequency of key terms an affect based time series can be constructed, allowing for an evaluation of changes in polarity within a given corpus over time. Following is an outline of the steps through biologically inspired algorithms may be applied to the refinement of a given lexicon, provided some external metric for evaluating the performance of the analysis a given lexicon may be pruned, removing terms misclassified within the given domain.

- IMPROVEMENT OF REFINED LEXICA (CHAPTER 4) Presents an examination between sentiment extracted from a news corpus concerning the US economy and the future movements seen within three market indices; the DJIA, S&P and VIX. These indices are selected as they have been presented in previous works within sentiment analysis of financial news, providing a mean for comparison. Firstly an affect time series obtain from a number of *Legacy* lexica is evaluated, the improvement obtained by using the automatically refined lexica produced from this methodology is presented.

- CLOSING REMARKS (CHAPTER 5) Concludes this work and discusses future areas of research, such as the potential for an examination on the impact of news flow, based on the level of information arrival rather than its content and the impact of re-release of old news data. Secondly the potential for adapting this methodology and implementing it towards other areas of textual analysis not solely sentiment of financial news.

Chapter 2

Motivation and Literature Review

2.1 Motivation

News sites, blogs, forums and user generated content sites have allowed for the generation of an unprecedented quantity of textual information. As such the development of tools for the indexing, and organising of such data is a field of expanding interest both academically and commercially.

The goals within textual analysis vary from a wide range of aspirations, content analysis, text summarisation, text classification, spam identification and sentiment analysis are but a few. Examining human textual content posses a number of challenges towards the researcher, human language is filled with subtlety, terms must be considered not simply with regards to their surrounding content but also the context and domain to which they are being applied.

The motivation is to explore how to asses sentiment of traders within a market; this motivation arises from past price, returns, earnings or volume movements which cannot be explained by endogenous variables alone. One can look at sentiment analysis of a random sample of people in a specialist situation; finance [89] economics [44], Lasswell introduced sentiment analysis or rather sentiment proxies, sets of terms used to express a feeling or desire about people [55], abstract things such as equities [104] or notions expressed in consumer good reviews [110].

The information related to sentiment contained within a given set of words, labeled typically as positive, negative or neutral categories, comprising unstructured or semi-structured texts can be viewed as an information extraction task either performed automatically or manually [54]. This task is performed by experts and intelligence officials, for example, who present their analysis as structured information in terms of opportunities and threats respectively. As defined by modern literature this is motivated by the promise of statistical techniques that may aggregate frequency, and locational information about linguistic units for the purpose of

producing structured information from unstructured data. In this case the structured information is a sentiment proxy as indicated by the usage (frequency) figures for the so called negative or positive words in conjunction with a key topic (location).

In the above discussion the sets of words being referred to *Positive, Negative or neutral* the technical name for this set is a glossary; a *list with explanations of abstruse, antiquated, dialectal, or technical terms; a partial dictionary.*[1] There are intuitive reasons for building these dictionaries as outlined by Stone et al. [100] where the sentiment annotation was carried out by political scientists, linguists, psychologists and computer scientists for annotating a word with its sentiment category. This method of work was imported into financial sentiment analysis in its entirety by other workers [104].

The criticism for this whole sale import of a list of annotated words has lead other researchers [37,64] to claim that classification was not right or misleading. However the critics of old word lists have used their own word list in order to evaluate the impact of sentiment on prices and returns. It is possible that an annotated word list is a necessary prerequisite of sentiment analysis if indeed this task can be carried out by looking at words as sentiment proxies. The methodology proposed here is this, beginning with a selection of words, and by using an objective measurement of performance or error of each random selection and through selection mechanisms drive the word list towards the optimization and refinement of these given words. The recent developments in evolutionary computing [27] are found to be useful and will be used for the methods and techniques for a set of randomly selected words trial solution; to predict market returns. Section 2.2 begins with a discussion of the development of textual analysis and its movement towards content analysis allowing for the integration towards a number of varying academic fields. From this the issue of text analysis and the specific domain of sentiment analysis is examined.

Following the area of sentiment analysis is examined in section 2.3; a growing domain with the goal of from the usage of statistical and linguistical analysis extract the emotion or sentiment from a given document; extracting such qualitative information towards the development of quantitative information. This is followed by a discussion on the growing field of sentiment analysis on financial information, aiming to extract sentiment information and examine the impact it has on financial market behaviour.

In section 2.4 flow of news information, the variations of quantity of news arrival over time is discussed. Of specific interest is the issue of *collisions*, news items re-released multiple times over time. Such re-released may be either exact in nature where there has been no alteration to the news items content, or with minor alterations within the news item's content. The impact such news items would have on any textual analysis and the means for the identification of such duplicates are topics for examination. Section 2.5 examines the area of evolutionary

[1] Oxford English Dictionary.

computing, an area where algorithms model certain behaviours seen in nature. Two algorithms are presented genetic algorithms and particle swarm optimization, both have been shown to perform well in certain problem areas. The aim of this work is to combine these algorithms with a glossary based approach to sentiment analysis of financial news. Closing is a brief discussion regarding the key topics which have been outlined within this chapter.

2.2 Content Analysis

> *There is the operational problem of analysing the semantic content of messages; this step has come to be known as content analysis.*- J. B. Carroll Study of Lang. iv. 120 OED

Content analysis aims towards the extraction of information from some form of human communication within the social sciences. The aim is to put forward methods by which researchers may identify some means to extract meaningful information from communications [94]; books, articles, speeches or any form of human communication and allow for the quantification of such qualitative information. Harold Lasswell put forward simply *"who says what, to whom, how, and with what effect?"* [54], with Holsti proposing the additional question of "why" [40]. Through the application of grounded scientific methodologies and statistical analysis unstructured information may become structured.

Early work within content analysis was aimed at the manual linguistic examination of human texts, term frequencies were manually counted and reference lexica were compiled from large number of texts across numerous fields and topics. Such resources proved invaluable as they provided a systematic approach to the examination of terms within texts. Such reference lexica would provide researchers with an objectively compiled dataset which records how often a given term occurs across texts. It may be concluded that if the frequency of an individual term within a reference text is significantly greater than that found across a wide range of texts that such term should provide some valuable information within the texts domain, view or authorship style, as the repetition and frequent usage of a term within a language indicates its acceptance.

Much of this early work was aimed at an examination of political texts, examining the news coverage regarding elections in an attempt to gauge support for individual candidates [55, 75], while others aimed to examine the political stances of candidates through an examination of terminology used within speeches [56, 57, 58]. Indeed much of Lasswell's early work was aimed at an examination of language within political texts and particularly propaganda [53].

While a number of news sources may not inherently endorse candidates within elections studies have noted that on occasions a single candidate may receive a significantly disproportionate level of coverage within a sources news coverage [3, 11] having a potential bearing

on the outcome of an election. Early attempts of such information extraction relied on an examination of the amount of space with a news source given to a particular candidate or party.

2.2.1 The GI Dictionary

Pioneering work by Harold Lasswell in the early 20^{th} century has used sentiment to convey the idea of an attitude permeated by feeling rather than the undirected feeling itself. Lasswell's work examining the linguistical content in political and economic texts was further developed with later works by Philip Stone creation of the Harvard Dictionary of Affect or the General Inquirer(GI) Dictionary [2]. The General Inquirer dictionary has over 11,000 entries in it and the entries are in many categories, there are 25 categories where each is divided into sub-categories making a total of 82. The most important is based on Osgood's semantic differentials that comprise adjectives that express evaluation(words covering negative and positive affect) potency (words indicating strength and weakness of an event or opinion) and activity (words expressing active or passive attitudes). The evaluated words comprise a third of the entries, the remaining categories are important with words specific discourses (political or military). The other attributes are mentioned in table below in table 2.2. The point here is that each word may have the attributes of one or more category of the GI dictionary. So the military term *"warfare"* has a negative evaluating, is strong in potency and shows activity, while *"able"* is positive, with weak potency and passive attitude.

Terms may be tagged an number of times across the original categories, as illustrated in figure 2.1 a significant number of terms are tagged across multiple dimensions, with as much as 10% being tagged across 8 categories, and a single term across as many as 9. Domain topics range from *Political, Economic & Military* and similarly may contain a emotional, strength and activity tag assigned to them. Such as the term *Advantage* termed as being a strong, positive economic term.

Table 2.1: Distribution of diametrically opposed terms in GI

	Polarity			Ratio
+	**n**	**-**	**n**	
Positive	1915	Negative	2291	1.19
Active	2045	Passive	911	0.44
Strong	1902	Weakness	755	0.39

The distribution of terms across the above cited categories is summarised in table 2.1; there are significant differences in the number of terms across the categories, while the ratio of *Positive to Negative* 1 : 1.19 across the second and third axis there is a significantly greater number of terms across the positive axis than negative. So as to deal with multiple

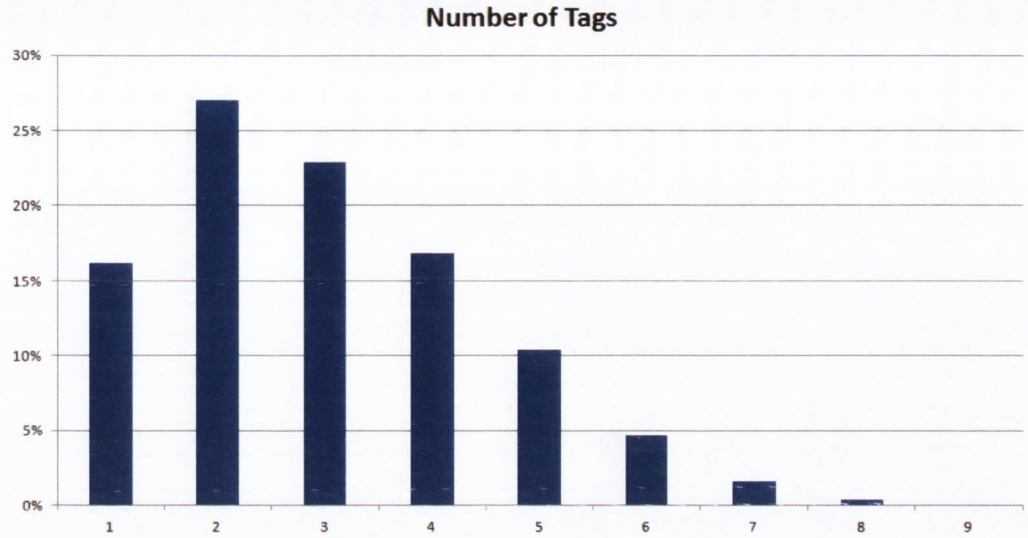


Figure 2.1: Percentage of Terms by Number of Tags GI

interpretations of individual terms multiple entries are possible, accounting for nouns, verbs, adjectives and other part of speech variations. As shown in figure 2.2, a significant number of terms have multiple entries within the General Inquirer, though the usage of linguistical analysis it is possible to identify which variation of the term is being used within a given sentence, which should allow for the correct semantic identification of an individual term.

A sample of terms contained within the negative and positive affect categories of the GI are presented in table 2.2. As can be seen the dictionary covers a wide range of terms, some of which such as *commoner* and *abdicate* are somewhat archaic and would not be expected to be seen commonly within modern political or economic texts.

Table 2.2: Selection of Positive and Negative affect terms within the GI.

Positive	Negative
hero, excellence, trophy	abdicate, decadent, default
experience, playful, aspire	alien, deficit, loss
unforgettable, nominate, super	beggar, exile, tax
faith, reconciliation, ascribe	break, hideous, horrify
amour, righteous, interested	brutish, disappointment, abate
realistically, lover, share	capital, commoner, fearsome

While such dictionaries are a valuable resource for textual analysis, it may not be applied directly towards analysis without further consideration and refinement based on a human knowledge of the language of the given domain. Studies have demonstrated the value of

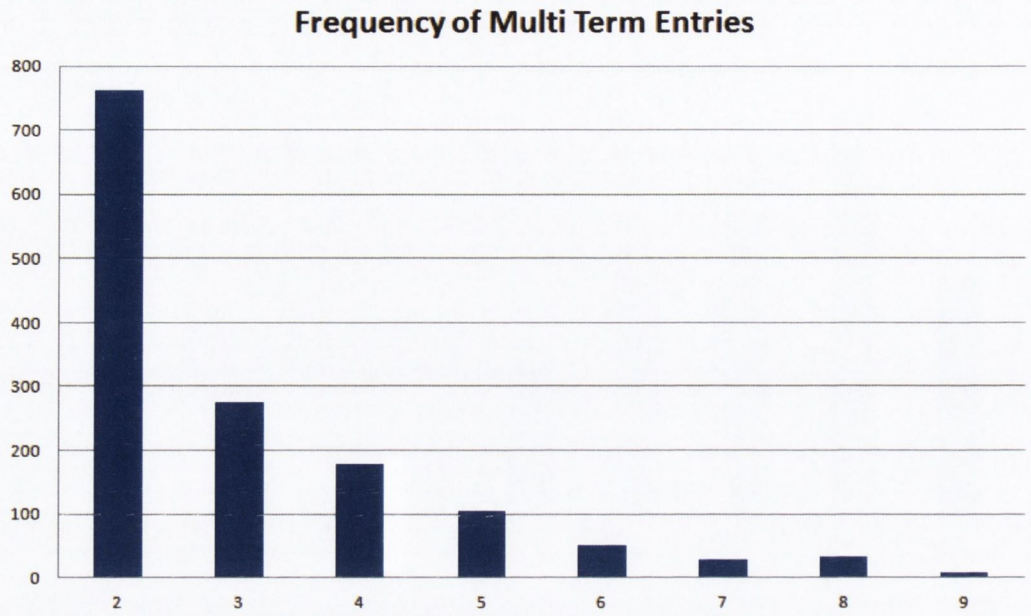


Figure 2.2: Frequency of Multi Entries within GI

custom lexica based on the domain in question for text analysis and text classification [107], yet note that such work may be impractical due to the requirement for an individual lexicon for every domain of interest. The development of domain specific lexica is an arduous and time consuming task, while for film or product reviews such work may be carried out by any individual, highly specific areas would require extensive work with experts within the field, legal, financial or political analysis would cover a wide range of terms and concepts not necessarily clear to the general public.

2.2.2 Variants of the GI

As noted when examining textual analysis from a lexicon based approach the decision of an appropriate lexicon source is of key importance. While comprehensive lexica such as the General Inquirer provide a source which has been compiled through large scale human intervention, providing classification of terms across a wide number of categories/domains issues continue to arise. Within textual analysis the domain of the application must always be considered; reviews for products, movies or services are clearly different in that while the aim of these texts is to serve the same purpose namely a recommendation or critique each possesses its own specific linguistical style [110, 117]. Paul Tetlock used the GI dictionary as was constructed by Philip Stone and relied primarily on the evaluative word list to estimate the sentiment. It is not clear from Tetlock's work [103, 104, 105] that he looked at the domain specific uses of certain words. For example the word *share* has a positive connotation of

meaning in the GI, *competition* a negative one and *pound* is given the negative evaluation strong potency and activity. If care is not exercised in disambiguating the use of these words then it's a possibility that the affect count would increase. More importantly when new terms are introduced then used to convey negative affect (Sub-Prime mortgage) the one would underestimate negative affect in those texts. So recently researchers in finance have used variants of the GI dictionary in two possible ways first manually updating the GI dictionary and "fixing" the polarity of words. The second alternative is to use machine learning techniques to identify words that are more frequent when the markets are rising and words which are more frequent when markets are falling [51].

Loughran & McDonald

The usage of the GI dictionary recently by economists and financial researchers for the usage of examining textual content within financial texts in recent years is a testement to the comprehensive and value of the GI developed by Stone [104].

However it is only in recent years that the ability to rely on the GI as it is presented in its original form has been questioned. Loughran & McDonald [64] provide a discussion regarding the potential flaws found within the GI when examined against financial terminology. They note terms such as *share* and *company* that; when applied to terminology within the financial domain have no emotional bearing So as to counter this they propose both a positive and negative term list based on their own knowledge of financial terminology. Firstly they examine all terms which occur in at least 5% of document within a corpus of financial reports; these terms are deemed to be of relevance in financial language. Following they consider each term and assign appropriate terms to a *financial negative* or *financial positive* category. This produces a negative word list of 2,337 terms which they propose are accurately classified when applied to financial news, and a positive financial word list of 353 terms.

Through the usage of such custom lexica they develop affect time series which they regress against movements of share prices for a given company. This is contrasted with a modified version of the negative and positive word list within the GI whereby they include inflections of terms. They find that the usage of custom dictionaries produce different results when compared to those obtained from the usage of the GI. The authors go on to claim that the usage of their dictionaries in identifying affect laden words, when used in a regression equation for estimating returns appear to be closer to observed results when compared to an affect time-series produced using the original GI.

Da, Engelberg & Gao

Similarly; Da et al. [16] consider the flaws of the GI when examining financial language. They proceed to use the dictionaries developed by Loughran & McDonald, and contrast the results obtained from the usage of the original GI dictionary. They examine the frequency of such affect terms within financial news regarding individual companies and examine such frequencies against future market movements. While the used *Financial Negative* & *Financial Positive* were developed for the examination of somewhat more technical reports they again note superior results when predicting future market movements.

Machine Learning

An alternative approach to lexicon development in financial sentiment analysis is through the usage of machine learning techniques. Antweiler and Frank [5] examined the relationship between the frequency of negative and positive affect terms within financial message boards and news items against financial markets. Rather than relying on a given lexicon here term lists were generated through the usage of a Naive Bayes classifier. A training corpus was developed containing 1,000 messages retrieved from financial message boards, each was assigned to one of three categories *Buy, Hold & Sell.* These authors do not relate buy, sell and hold directly to positive or negative content of message board postings. The buy signal for contrarian investors relates to the market bottoming out and dominance of the bearish sentiment. Of issue here is the requirement for a number of human annotated messages for the purpose of training the system. While this may appear a simple but tedious task, it is not without the possibility of errors. As noted by Pang et al. [81] where when developing a word list for film classification terms provided through human interview performed rather poorly. Pang et al. propose that a corpus based approach my be superior than relying on prior intuitions. This work aims to combat the need for human annotation in a manner similar to Koppel and Shtrimberg [51] where movements in the market can be used for training purposes. If there is a relationship between sentiment within financial news and next day market movement it should be possible to use the market movements as a form of tagging, where an upward movement is *positive* and a downward movement is *negative*. Antweiler and Frank lays the foundation of what appears to be machine learning techniques discussed in section 3.3.

2.2.3 Discussion

This section has discussed the area of content analysis and in particular the usage of Lexica within the field of textual analysis. Stone's General Inquirer was an innovative approach towards the classification of terms regarding their polarity, domain, potency and activity has

provided researchers with a tool allowing for extraction of textual information from a given text.

This work aims to develop a means to examine the issues of contemporary and legacy lexica, as while comprehensive legacy lexicon prove a valuable tool towards researchers they are not without their flaws. In each instance such lexica must consider the domain to which the corpus of interest refers to, language is dependent on the target domain whereby the affect meaning of a term within one domain or topic does not hold true for others. And while the development of expert lexica have shown to be advantageous they remain flawed in their reliance on large scale human development and agreement requiring possible in-depth knowledge of the domain in question. This is achieved through two machine learning algorithms Genetic Algorithms and Particle Swarm Optimization. Both methods are developed with the goal of taking a legacy lexicon and through feature selection identifying terms which are of interest in textual analysis. The basis behind Genetic Algorithms and Particle Swarm Optimization is presented in section 2.5.

Such a methodology should improve the performance of any analytical evaluation yet also provide information regarding the linguistical style of the given domain. Providing a lexicon which having been refined identifies terminology which is correctly classified and that which is incorrectly classified.

2.3 Sentiment Analysis

Sentiment analysis has been defined as an attempt to determine the attitude of a speaker or writer of a document [100, 104]. The history of sentiment analysis can be traced back to the emergence of content analysis where documents were first encoded manually in terms of their polarity, potency, and activity for example and then analysed by machines [76]. A basic tenant of sentiment analysis is that one can determine the affect of the authors or authors by doing a quantitative analysis of the documents written by the author or authors. This may involve for instance estimating the news flow, for example the number of messages posted on a site, or a frequency count of given polarity categories [104].

Sentiment analysis begins with the premise that texts may contain an emotional view or argument expressed by the author. This is followed by the concept that the aim of the text is such that the author(s) may wish to pass this view onto the reader; conveying information, knowledge or attempting to influence the readers' views on a given topic. If a text written by an individual attempts to convey an emotional quality we may suppose that within this text through the selection of terms by the author sentiment has been *encoded* within the text. Lastly that, through the application of machine learning techniques and statistical analysis of document contents and structure we may extract such information and identify the *polarity* of

a text, quantifying such qualitative data regardless of the topic domain, producing in a sense a proxy to the sentiment which the author wished to express.

Other works may be concerned more with the classification of texts according to more complex human emotions, such as fear, hatred, anger or depression [19]. Works have also examined the ability of systems to automatically identify the objective or subjective nature of a text.

Sentiment analysis can also be considered as a subset of information extraction and text classification; while text classification aims to determine the topic towards which a given text is in reference to, within sentiment analysis we aim to determine the polarity of a given text. While various textual analysis classification techniques aim towards a binary classification system such that document D may be assigned to category C with a certain level of probability, within sentiment analysis we may not view classification as binary but rather as a level of the negativity or positivity of the text. Considering sentiment analysis of film reviews, a film is not necessarily *Good* or *Bad* but generally assigned a score on a scale representing exactly how good or bad, or neutral, it may be viewed.

With the growing importance of online-shopping, product reviews have become a key area of interest both in terms of commercial and academic importance. Sites such as Amazon & TripAdvisor allow for the generation of product and locational reviews provided by site users, given the impact of such reviews towards future customers purchasing patterns such review content is a key are of interest and importance. The extraction of sentiment from reviews is difficult as a single review may express multiple opinions of differing features on a single product. users to provide personal reviews on items and venues. If examined on a sub-sentence level analytical techniques may be able to extract the varying views on individual features, such as an aesthetically pleasing design yet lacking in software for a computer product [117].

As with other areas of sentiment extraction the selection of an appropriate lexicon, techniques have been implemented so as to develop a domain independent reference source for product reviews. Where as movie reviews language may be complex and requiring a domain specific dictionary [107] due to the possibility of plot summary descriptions within the review such as describing the *horrible* nature of the villain, within product reviews authors are probable to use emotionally adjective terms purely as a description of their experience, such terms may be extracted from a complex lexicon source such as WordNet [42].

A further problem when examining the sentiment of product reviews is that reviews on sites such as Amazon tend to be vary in length, with several being extremely short while others providing a more comprehensive review. Due to the variations in length weighting methodology is of importance. While a short review would generally contain an high level of relative emotional/affect terms, a longer and more comprehensive review's value would be lower yet more useful to potential customers in their decision making [50]. The marketing

industry appears to be keen on using customer reviews whether they are about books, full length texts or about film reviews which are essentially sentences or the Likes/Dislikes seen on social media. The quantification of the views of one writer, or indeed a whole cohort of writers, evaluates a product or a film is usually conducted on word frequency counts or co-word analytics [102]. There is an emergence of literature establishing a correlation of the content of the reviews and external variables.

Early work by Pang et al. [81] turned to an examination of sentiment analysis towards film reviews. Using a number of varying textual classification techniques they aimed to classify documents within a pre-annotated corpus as being positive or negative film reviews. Text classification techniques such as Naive Bayes and Support vector machines and Maximum Entropy are shown to perform well, in both unigram and bigrams analysis when looking at the frequency of presence of terms anywhere in the text. While the results returned are promising the authors note that such results are inferior when compared to those using similar techniques when applied to standard topic based classification. Such results posses the potentially more complex problem in regards to the automated extraction of sentiment contained in human speech and texts.

Tong [107] allowed for the generation of a sentiment time-series based on the frequency of affect terms, however as such terms were predefined by the author such a lexicon is domain specific, an expansion towards other fields would require further time consuming human decision making. Independently, Turney [110] proposed the examination of sentiment however across varying review topics[2]. Similarly while results were promising difficulties arise when examining film classification. A key issue of interest within this work is the author's commenting that when examining terms across varying domains it is necessary to consider varying interpretations of terms towards their topic domain. Within automobile reviews "unpredictable" would be a highly negative term yet when applied towards a film such a term may be viewed as positive. Work within sentiment analysis may concentrate on a document, sentence or sub-sentence level. There are number of advantages with regards to the examination of sentence and sub-sentence level analysis in terms of texts relating to product reviews; where users may praise single features while addressing flaws in others.

When research concentrate on the usage of a pre-defined lexicon for sentiment analysis issues arise from the negation factor within human language. Returning to film and product reviews the comments "Not Bad" or "Not Good" demonstrate the importance of control for term negation, while simple methods for controlling for this occurrences are available through simply labeling all negation terms as *NOT_* which do lead to improved results there remain other issues frequently within the field of varying language usages across varying domains [81]. Other studies have shown that despite controlling for the negation of terms an examination

[2]Banks, Films, holiday destinations and automobiles

of negative sentiment within documents may be of greater value. When examining Google queries [16] it is probable that individuals will not enter a term while seeking its negation, it would be counter-intuitive to speculate that a query for *Growth in Profits* would be entered when expecting results discussing a drop in profits. This may be due to the concept that human beings are more emotional or better influenced when presented with negative information, such that we positive information can be considered as the absence of negative information. Past experiments have moved further to demonstrate the differences in human responses when presented with similar facts yet phrased differently to lend importance towards positive outcomes when compared to negative outcomes.

2.3.1 Rational and Irrational Behaviour

Standard theories tell us that markets are rational entities, traders behave in a purely rational manner; considering purely factual data towards determining their trading strategy.

Within the Efficient Market Hypothesis produced by Eugene Fama as soon as information becomes available it will be absorbed into the market share price instantly. Market movements are the result of rational behaviour based on technical and quantitative information. All currently available information is already incorporated into the market price and as such would be absorbed before any media reporting may be published.

However it is easy to see how such behaviour is not seen within the real world, theory tells us that prices move in a random walk; the returns of prices when plotted will form a classic bell curve, with most price changes being of very small amounts, while large price changes are possible their occurrences should be so infrequently they need not be considered. Yet there is sufficient empirical data which shows that markets are not always rational, that these predictions regarding the normality of returns is far from true. Following given theories market bubbles and busts should not occur, and even if they do their frequency should be so rare that they may be considered one off events.

Mandelbrot has shown that the market crash within the US in August 1998 should never have happened; on August 4^{th} a drop of 3.5% in the Dow Jones Industrial Average was seen, three weeks later a drop of 4.4% on August 31^{st}. According to the efficient market hypothesis, based on normal distribution of returns confined to 4 standard deviations it should not have happened. The probability of the August 31^{st} was one in 20 million, while the cumulative probability of both occurring, based on financial theory is less than 10^{50} "odds so small they have no meaning"(Page 4) [67].

The cause of such events are unknown and unpredictable, however the human factor must be considered. Humans do not behave rationally, when evaluating outcomes individuals concentrate on the positive and downplay the likelihood of the negative. Studies show that

given two situations each containing identical probability of outcomes individuals concentrate on those they view as risk aversion.

2.3.2 The Role of Sentiment Analysis within Finance

Following is a discussion regarding the analysis of sentiment on texts in the field of sentiment analysis of financial documents. Works aimed at extracting sentiment from financial news stories [18, 90, 103, 104], firm reports [37, 64], Google queries [17], message boards [5] and Twitter feeds [9, 88] have in recent years been of growing interest to investors. Most research and applications are aimed at the development of systems which would allow for future predictions of financial returns or indications of future market volatility.

Varying theories persist regarding the role of sentiment towards financial instruments. Past works state that market behaviour is based purely on rational behaviour with all information being currently incorporated within the firms share price based purely on technical information such as profits and earnings. Holding to theories of a purely rational and efficient market such information should be meaningless, any technical information provided has already been incorporated into any current stock price. While any emotional or sentiment data within news will have no impact on a purely rational market. However as far back as 1936 noted economist John Maynard Keynes coined the term *Animal Spirits* viewing that economics is not entirely governed by rational behaviour and mathematical analysis but by the natural tendency for human optimism and actions governed and motivated purely by human emotions and feelings. Holding towards the prepositions regarding sentiment within texts outlined earlier it is desirable to examine if such emotional motivations may be derived from relevant texts, their impact on investor's extracted and quantified. If this theory hold to be true two following interpretations are possible; information contained within news items may simply be a barometer of a firm's "health" and while an examination of such information may show it to be consistent towards market behaviour there is no new information contained within the news item. Alternatively it has been proposed that news sentiment may in itself influence the perception of investors and as such impact on an investors' behaviour.

A case for consideration is that which occurred in 2008 when a news story regarding United Airlines' 2002, through a flaw in an article's time stamping a news story was picked up and indexed as being new and included on Google News. Individuals reading the story or simply the headline 'UAL Files for Bankruptcy'concluded that UAL was filing for bankruptcy for a second time and reacted accordingly. Following this item was posted on Bloomberg news services, within a number of hours UAL's share price had dropped 73%. While the error was corrected on news-services accompanied by a statement by UAL regarding the emergence of this old news story which followed a rise in stock price this error was not fully corrected for.

According to the efficient market hypothesis clarification regarding this incident would have been absorbed instantly into the financial system and the stock price would behave accordingly, if there had been no other relevant news items or information prices would correct quickly to the companys previous price. However studies examining the event have shown that while other airlines who's prices were affected by this event did indeed correct, UAL's share price was adversely affected for four days [68].

Bag of Words

Within textual analysis a commonly applied method is a Bag of Words(BOW) technique, which examines the frequency of terms independently of their position within a text and the surrounding content.

A vector is constructed where each element records the frequency of an individual term, while simple in its concept and implementation BOW methodologies have been shown successful within textual analysis and, in instances outperform methods of greater complexity [10, 81]. Alternatively a bag of words may be a binary vector indicating the presence or absence of specific terms within a dataset, such an instance may be termed *Feature Presence* (FP) , while a model which includes the frequency of the feature within the dataset as *Feature Frequency*. While on first consideration the recording of purely the presence or absence of a term appears simplistic researchers have produced successful results within textual analysis [81]; however other works have argued that the performance of such methods may not consistently outperform feature frequency [78].

Within a Bag Of Words technique documents may be summarised as

$$D_j = \{W_{1,i}, W_{2,i}, .., W_{n,i}\} \tag{2.1}$$

Where i is the frequency of term j in document D. Terms are generally examined in a unigram format independently of their surrounding vocabulary, while this may lead to the meaning of terms being altered by their surrounding vocabulary such as the case of the usage of a negation term developments in textual analysis allow to deal with such events. An expansion of this method allows for the examination of Bi-Gram terms where the frequency of terms commonly found within the surrounding vocabulary will be examined as a single unit of the document vector. An issue for consideration within a Bag of Words approach is the elimination of terms which in themselves provide no novel information towards textual analysis or classification. Terms could be removed according to human annotators however such work is time consuming and noise terms from one domain will not necessarily transfer effectively to other domains. A number of methods are presented by which document vectors may be refined, common method include the exclusion of all terms deemed as *Stop Words*; while there is no defined list

of *Stop Words* terms such as *"the, a, is, it.."* and other such function terms are contained within documents and speech generally for grammatical purposes and they themselves contain little affect bearing information.

In combination researches commonly apply a term weighting feature towards textual analysis, the resulting analysis will be based on the relative term frequency across a training corpus and documents to be evaluated [69, 77, 80]. Through the assignment of weights towards individual terms based on the frequency of their occurrences across a training set of documents.The impact of common terms which may have no bearing on the documents content will be assigned a weight which will mitigate their impact on further analysis. Further refinement method such as Weirdness [4] evaluation examine the frequency of a term within a text against a given reference corpus allowing for a methodology for the identification of important terms within the given domain. Alternatively researchers may wish to examine the frequency of predetermined terms. Such terms may have been assigned to a category either within general language of domain specific usages.

While it may appear intuitive to include stems of terms as their meaning and concept would contain similar information with variations in terms begin for purely grammatical purposes, previous works [59, 69] have demonstrated that such inclusion may actually be detrimental to classification results while increasing the computational time of any such training methods.

Weirdness

Weirdness allows for an automated means for the extraction of key terms with a given corpus based on their relative frequency within texts when compared to a reference corpus. By comparing the frequency of terms across a specialist corpus these frequencies may be compared to the terms relative frequency found within a non-specialist corpus. By selecting a non-specialist corpus such as the British National Corpus (BNC); being compiled from 100 million term across a wide range of topic domains it is possible to estimate the average frequency of a term within the English language. If a term occurs significantly more often within a specialist corpus when compared to a general language corpus it may be concluded that this term may be of interest within the specialist domain providing possible information. Calculating the *weirdness* of a term is based on an examination of the term's frequency within the specialist corpus to the non-specialist corpus according to equation 2.2 [4].

$$Weirdness = \frac{W_s}{t_s} * \frac{t_g}{W_g} \tag{2.2}$$

Where W_s and W_g are the frequency of the term within the specialist corpus and general language corpus respectively, while t_s and t_g are total number of terms within the corpora. In this case Ahmed et al. have introduced the notion of z-Scores, which normalise values

according to the standard deviation and mean across the data set, for both relative frequency and weirdness. By evaluating the weirdness value, if a term's weirdness value is greater than 1 it is occurring more frequently within the specialist corpus than would be found randomly according to the general language corpus.

If this weirdness measurement is applied to film reviews or scientific documents it would be expected that a number of terms to have a high weirdness value as they are at times domain specific, this is certainly true when examining scientific documents as many of the terms found within are domain specific [33]. With a threshold value determined the resulting analysis allows for the extraction of a lexicon specific to the given domain; whereby any term with a weirdness value greater than θ can be considered informative when attempting to examine texts within the field. Where θ is a predetermined threshold value. The weirdness of closed class words is typically of the order of unity, showing that closed class words of a given language are distributed similarly across text types. However the open class words especially the domain specific words are generally greater than unity.

Term-Weighting

When looking at the frequency of individual terms or n-grams within a corpus for the purpose of textual analysis, rather than counting simply term frequency the application of a term weighting method is a common approach for the identification of key terms.

Through the implementation of a weighting function *noisy* or commonly occurring terms' impact on any future analysis should be mitigated as the weight of these terms will be decreased. If we wish to examine a corpus regarding financial or political events we may expect number of terms to occur more frequently that within general language texts. If we wish to classify a number of texts towards their topic and we identify the term *Economy*, this would be a strong indicator of the documents topic, however if examining a specialist corpus such a term provides little to no information. While a number of weighting methods are available the most common implementation is a term frequency inverse document frequency method ($tf-idf$) , which aims to quantify the rareness or importance of a term within a given document. If a term has a high frequency within a given text but a lower frequency across the corpus as a whole it may be concluded that such a term has valuable information towards the given text. Firstly the inverse document frequency of a term (idf) is calculated according to the formulate 2.3.

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{2.3}$$

Where $|D|$ refers to the total number of documents within the given corpus and $|\{d \in D : t \in d\}|$ is the number of documents within the corpus containing the term t, if the term occurs in 0 documents within the corpus a zero may be assigned or it is possible to inflate the frequency of

all terms by 1.

Using equation 2.3 $tf - idf$ may be calculated according to equation 2.4:

$$tf * idf(t,d,D) = tf(t,d) * idf(t,D) \tag{2.4}$$

With $tf(t,d)$ being the frequency of term t in document d. Returning to the commonly studied field of movie reviews it may formulate that if term t is found very rarely within pre-annotated positive reviews yet frequency within negative reviews that such a term has a high probability of being negative within the confines of movie reviews, as this repetition of a term indicates its acceptance within the target domain.

2.3.3 A synthesis of work in financial sentiment analysis

The literature on finance analysis to report use of different kinds of lexica, variants of GI and machine learning dictionaries; different kinds of texts have been analysed for identify the mood of the market; and the impact financial news has been established on price movements; company returns, proxy while using a variety of different environmental models. Four have been selected here

Table 2.3: Comparison studies of sentiment in financial news.
This table summaries four studies within the field of sentiment analysis of financial text and contrasts them with my own research. Each study applied a bag of words technique to extract sentiment from the given source and carried out regression analysis with the aim of improving prediction of market returns.

	Antweiler & Frank [5] (2005)	Tetlock [104] (2007)	Loughran & McDonald [64] (2011)	Da et al. [16] (2011)	My Proposal
Lexicon	Machine Learning Annotated Texts	General Inquirer	General Inquirer Refined through Expert Knowledge	General Inquirer Refined through Expert Knowledge	General Inquirer Refined Through Machine Learning
Texts	Financial Message Boards & News Items	WSJ Column	Finance Reports	Google Search Trends & News Items	Financial News
Time Series	Company Returns, Trading Volume & Volatility	DJIA Returns	Company Returns, Trading Volume & Volatility	Company Returns, Trading Volume & Volatility	Index Returns
Output	Excess Returns, Volume, Volatility	Excess Returns	Excess Returns	Excess Returns, Volume, Volatility	Returns

Four past studies examining sentiment analysis within the domain of finance are summarised in table 2.3, note that researchers have begun with the usage of the General Inquirer lexicon [100] for the purpose of sentiment analysis, following refining the lexicon through their own knowledge within the domain [64]. Loughran & McDonald comment that the GI lexicon is advantageous as it has been produced in a systematic manner placing it beyond the control of the researcher, however they then proceed to generate a lexicon based on the authors' subjective opinion on financial language arguing that when examined within the confines of financial

terminology as much as 73.8% of terms are not negative within the financial domain [64]. The authors present terms such as *capital, board, liability, foreign* and *vice* as notable examples. Within these studies researchers expand upon the given lexicon to develop through manual evaluation a domain specific lexicon. In each instant the author(s) construct an affect based time series, this time series is constructed by counting the frequency of a list of key affect terms and a sentiment score is generated for each datapoint. With this time series a regression model is constructed which aims to demonstrate a relationship between sentiment and future market behaviour.

The lexicon of financial terms produced by Loughran & McDonald was later used by Da et al. [16] when examination sentiment within firm specific news items contained within the Dow Jones Newswire articles. While the proposed lexica used were developed for the usage against company report fillings rather than financial news articles it may be possible that while less technical the variation in language would be minimal, such that if a term is negative within a financial report this would hold true within news reporting. This affect time series was evaluated alongside a Google search volume index of individual companies, the researchers examined the frequency at which people searched for a specific company, so as to ensure accuracy and concentrate on examining traders search trends they use individual company market codes than company names.

Alternatively is the usage of machine learning techniques for the development of lexica; through the manual evaluation of text items Antweiler and Frank [5] tagged each as positive or negative in nature; from which through the application of machine learning techniques such as Naive Bayes a lexicon for a given corpus may be produced. This was combined with an examination of affect term frequency within financial news in the Wall Street Journal on firm specific news items. It should be noted that while Antweiler and Frank selected *LexisNexis* as their source for retrieving news items there is no mention of the possibility for collisions within their corpus and the potential impact this may have within any textual analysis.

Experimental evidence demonstrates the benefit of such domain specific lexicon; while a corpus of inter-company financial reports and fillings is notably different from that compose of financial news reporting [64], the target domain is comparable. While a lexicon aimed at more technical forms of reports may contain a number of terms rarely used within financial news reporting it is unlikely that the meaning of the terms is significantly different as the domain remains similar.

Through an examination of the frequency of key affect terms regression models may be developed to examine movements against a given financial time series. The resultant models include technical data such as market price changes, volume traded and other technical financial variables. The models will generally incorporate a lagged value of sentiment, this is so as to allow for sentiment information to be assimilated while also to ensure that no future knowledge

is used within analysis. Added to these is the inclusion of a form of measurement for sentiment; this inclusions of sentiment allows for an inclusion of *qualitative* data against a technical *quantitative* system. While the impact of sentiment has been shown to be to a far lesser extent than that of market fundamentals the inclusion within models of future financial movements has be shown to be statistically significant.

Tetlock [104] provides an examination of affect term frequency within a single daily column within the WallStreet Journal against the Dow Jones Industrial Average. This column *Abreast of the Market* provides a discussion on the movements and events observed within the financial markets during the course of the day. His selection of news source is of note as the DJIA was created by Wall Street Journal editor and Dow Jones & Company co-founder Charles Dow, this selection may be based on the assumption that given this relationship a significant portion will be devoted towards a discussion on the DJIA. Also rather than considering the collection of a wide number of news sources, a single news item per day is selected, this limits the work as it prevents and examination on the role of news flow and the quantity of individual news items being released on a daily basis.

Two methods for the measurement of sentiment are proposed; firstly an affect time series is obtained through an examination of the frequency of terms within two categories within the GI *Negative* and *Weak*. A secondary method producing a *pessimism* factor is proposed; firstly all terms across 77 categories within the General Inquirer are counted, from this through component factor analysis which collapses all categories into a single media factor which captures the maximum variance across all categories. By examining those categories which contain the greatest level of variance across a training corpus with the aim of extracting the most important semantic components and reduce redundant categories. An issue which does not appear to be discussed is the occurrences of terms across multiple categories, as a significant number of terms are classified across multiple categories with one across as many as 9 categories (figure 2.1) this may lead to double (or as many as instance 9 times) counting where initially all terms are given equal weighting, such an impact would inflate the importance of a certain terms. This multi variant analysis is in contrast to other studies where sentiment is composed of a single set of terms of equal weighting. The issue of multiple categories comes because a dictionary such as the general inquirer was created manually by psychologists, political scientists and economists headed by a native speaker of English Philip Stone. Combining the knowledge of subject domains and fluency in English assigned categories to which a word can belong, essentially it was the judgment of the GI team involved in the making of the GI lexicon. For me this was a point of departure and my attempt to reduce subjectivity. In the work reported in this thesis, categories are not assigned per-se; but make inferences about the impact of the use of certain words on the market movements. Having produced three affect time series a regression model is proposed where lagged values

of sentiment are examined against future market movements. Lagged values are used firstly to discount the possibility of future knowledge, secondly for the examination if sentiment is a short lived impact namely that given time once this information has been absorbed will market behaviour returns to fundamentals. This model is further developed through the inclusion of a number of exogenous variables; such as a detrended squared residual of the Dow Jones Industrial Average as a means to account for high levels of volatility while also introducing control variables so as to account for market crashes.

The results obtained within this study find a statistical significant relationship between Tetlock's produced *pessimism* factor and market returns on the following day. While sentiment does have a statistically significant impact on market movements on the following day further analysis demonstrates that this is followed by a reversal, whereby the impact on movements is reversed and the impact of sentiment is lost during the course of the week. This view that sentiment is short lived and that the impact of it is corrected at a later point is similar to examinations of the impact of abnormal trading volumes where given time markets revert to fundamentals [12].

2.3.4 Discussion

This section has discussed the area of sentiment analysis within the field textual analysis. Sentiment analysis' main premises stem from the argument that a given piece of text may contain an emotional or sentiment being expressed. From this it may be suppose that this sentiment which the author may wish to express in this text and is expressed through the selection of key terms by the author of the text. It is possible that the author was reacting to, or reflecting upon, an event which persuaded him or her to use the specific terms. If indeed such key terms express the sentiment it is possible that through an examination of the occurrences of such terms the sentiment being expressed may be extracted. This may allow for the quantification of such qualitative information allowing for an examination in a statistical manner. Within sentiment analysis the domain towards which the text is being applied is of importance. The interpretation of terms vary according to which topic or domain they are being applied. This work aims to examine such techniques in the domain of sentiment analysis within financial news. While financial markets are viewed to behave rationally; based on technical data and statistical analysis of market fundamentals such views have been brought into doubt. This work examines sentiment contained within financial news and the potential impact such news has on future market behaviour. Studies have demonstrated that through an analysis of the frequency of affect terms within relevant news items and other textual sources a statistically significant link can be observed towards future market returns.

2.4 Collisions and News Flow

When examining news items about a given event it is common for individual news sources to simply reprint an item produced from an alternative news source, AP NewsWire, Bloomberg, Reuters. If such texts are exact duplicates there is no reason to believe that these items contain novel information, the identification and exclusion of such text is a trivial matter yet if not considered may have a significant bearing on the analysis. However it is possible that the simple re-printing of such items indicate their importance within a corpus, a reprinted review may reflect that of a notable and therefore influential author, similarly if numerous sources reprint an identical news item this may simply be due to the importance of the event and reliability of the given source. Of further consideration is the issue of partial re-prints; within blogs or forums the quoting of previous posts would be rather common and while any resulting analysis would have already been included within the system it is worthy of considering that similarly to the reprinting of review this post may represent an important view within the current discussion. Within news events these may occur for a simple factual or grammatical clarification yet may also indicate the repeated discussion relying heavily on previous texts and being released for the purpose of clarification towards events as they unfold.

Through various means it is possible to quantify the textual similarity between two texts, allowing for the tracking and changing of a topic, event or review over time. It is assumed that instead of matching two documents word for word it is possible to use the notion of cohesion to create a representation of text. There are two types of cohesive devices available to a writer, lexical cohesion and syntactical cohesion. Where syntactical cohesion would be the use of determiners, when talking about sentiment in text we introduce the topic by stating we are investigating the presence of sentiment in text and subsequently suggestion that later on in the text negative and positive sentiment the determina *the* makes a definite reference to the topic being considered and links the two sentences. On the same level we can say Paul Tetlock devised a scheme and then later on suggest "he" concluded, linking the two sentences. Lexical cohesion focuses on open class words and links are formed in the text by the repetition of domain specific words, a so called cohesion graph can be used to compare the similarities of two different texts by comparing their cohesion graphs [21]. Rather than attempting to match documents on a structural basis seeing if the documents are near word for word identical in their ordering, it is possible to examine the frequency of certain terms. Kilgarriff counts the frequency of the n most common terms within two corpora, the frequency of the terms are compared against the expected frequency if both corpora were random samples [49]. By using a Chi-Squared test (χ^2) Kilgarriff was able to compare the difference between the expected and actual occurrence of a term within a document. Where if the size of corpora 1 and 2 are N_1 and

N_2, and the word w has a frequency of $O_{w,1}$ and $O_{w,2}$ the expected value $e_{w,1}$ is calculated as

$$e_{w,1} = \frac{N_1 * (O_{w,1} + O_{w,2})}{N_1 + N_2} \tag{2.5}$$

with $e_{w,2}$ calculated similarly. χ^2 can be calculated by examining the differences between the expected and the observed frequencies.

$$\chi^2 = \sum frac(O-e)^2 e \tag{2.6}$$

By comparing the difference between how frequently a term occurs and how frequently it would be expected to occur it is possible to gauge how similar the two text are. This approach has the advantage that rather than examining the structural content of a document the frequencies, which are far easier to calculate, are examined.

It has been proposed that while old news events have a lesser impact on market behaviour [105] when compared to new events and commentaries this information is non-trivial. The possibility that that the mere act of reprinting, correcting or expanding on a piece is in itself a reflection towards the importance of the original text(s) should not be excluded. While the identification of identical duplicates are trivial determining partial duplicates poses a challenge, if grammatical variations may be found at any point within the text, the alteration of a single letter would alter any simple examination of the proposed texts. Methods towards the development and methods for the identification of such duplicates will be discussed in section 2.4.2.

2.4.1 Text Sanitisation

Frequently documents retrieved from online repositories will contain a number of tags, the purpose of which is to classify and record the topics of the document while providing further information such as the author, origin and date. This information is of value however care must be taken to ensure that none of this *meta data* is included within the later analysis. A sample of an article retrieved from the LexisNexis online repository is presented in figure 2.3, this repository provides the ability to search across a large number of sources based on criteria regarding topic, key words and date.

Files may be retrieved in batches; however for some of the research it is of value to *split* such files into a corpus of individual text files. The consistency of the output format allows for the identification of the beginning and end of each individual file, while also extracting date information and removing all meta data. While such meta-data are of use as they index a number of key topics regarding the text such as the topics, persons mentioned or information regarding geographical location of the even when examining term frequency such tags may

```
            1 of 1122 DOCUMENTS
               The New York Times
      December 31, 1996, Tuesday, Late Edition - Final
Home Resales Surged a Surprising 1.8% in November
BYLINE: By ROBERT D. HERSHEY Jr.
SECTION:Business/Financial Desk
LENGTH: 671 words
DATELINE: WASHINGTON, Dec. 30

Spurred by falling mortgage rates, sales of existing
homes rebounded smartly in November[...]

LOAD-DATE: December 31, 1996
LANGUAGE: ENGLISH
GRAPHIC: Graph: "Leading Indicators" tracks index[...]
Table: "Leading Indicators: Component Analysis"[...]
Copyright 1996 The New York Times Company
```

Figure 2.3: A Sample LexisNexis Article Format.

have a detrimental effect on later analysis.

The beginning line always indicates the placement of the article within the downloaded batch and allows us to determine the beginning of a new document, as stated this provides

a consistent format of *Number* of *Number* Documents, this patter is easily identifiable through the usage of a regular expression statement: *"([0-9]1,) of ([0-9]1,) DOCUMENTS"*.

This is then followed with a line indicating the source of the document with the date following subsequently in a natural language format. In this instance it is simply required to extract the first term as the month in natural language, followed by digits indication the year of publication. From this comes the ability to extract articles in an individual format while indexing them in a machine readable manner. Given the aim is to track variations of sentiment over time towards the construction of an affect based time series the publication date of each news item must be stored. Unfortunately there is no guarantee that an exact timestamp of an article's release is included within the news item, as such articles are indexed and stored in a daily manner. For this purpose all articles once split are stored in a custom naming convention:
MM_DD_YYY – DOW@Headline.txt
Where *DOW* stores the day of the week, within this manner all articles can be stored in a simple machine readable fashion.

In regards to included meta data there is frequently a number of lines which provide information regarding the author and domain yet through manual examination of the files it is shown with consistency that the first paragraph marks the beginning of the documents

body. Following the document's body is a number of pieces of meta data, while there is no consistency regarding which tags are included there is a limited number and such once any of these manually identified tags such as *LOAD-DATE* or *LANGUAGE* the parser has reached the end of an individual article.

In this manner a corpus may be collected in a simple and efficient manner, little human intervention is required beyond the selection of document requirements, while all meta data is automatically removed and the document's body may be extracted and stored with the publication date returned. This automation not only allows for the rapid development of a large scale corpus but also ensures that the generation is carried out in a consistent manner.

2.4.2 Collision Detection

While examining the produced corpus for the purpose of experiments a certain issue was found where there appeared to be a significant number of documents with identical or similar headlines. On closer examination it was found that the corpus contained a number of articles which were either exact duplicates of previously returned news items or simply minor variations or expansions of others. These duplicates will be referred to as collisions; it was found that the inclusion of such duplicates may have a major bearing on any textual analysis output, if examining the impact of the reporting of an event these collisions could have a detrimental impact on any analysis. As these collisions represent either exact duplicates of previously published news items or minor variations of others it is possible that such articles may provide no new information in their own right. These collisions may simply be a mistake in the indexing of the news repository or were found to frequently occur simply for the correction of a minor or grammatical error and as such would provide readers with no new information and would be assumed to have no further impact on an individual's reaction to an event or piece of information.

However others were found to on occasion be an expansion of previously written pieces, adding paragraphs at some point within the document, generally in the interest of clarifying some point or expanded as events unfold and further information becomes clear regarding an event being discussed. The development of a methodology towards the identification of duplicate items was in itself a non trivial problem, while the task to identify an exact duplicate is a simple comparison operator others proved more challenging. Four individual methods for duplication detection were developed, each tailored towards a specific form of collisions for detection. Once news items have been collected each article is examined against all other articles released within a -1 day +1 day window, while there is no guarantee that collisions will occur outside this window when expanded the number of comparisons grows rapidly eventually reaching an infeasible number. Additionally a test examination whereby all articles

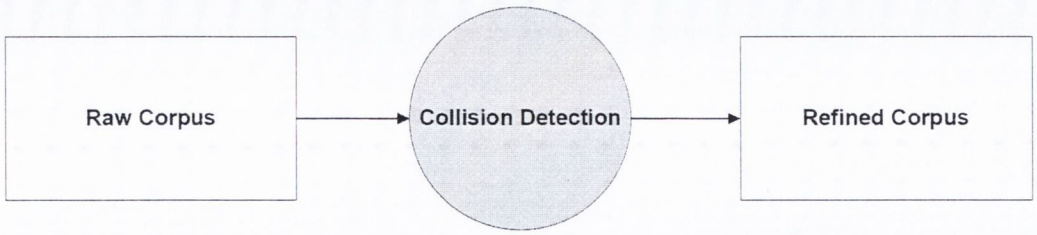


Figure 2.4: Corpus Refinement.

were examined against all other articles within the same month identified one single extra collision outside of this range over the course of a corpus comprised of a year's worth of news articles.

Exact Duplicate

The identification of instances of exact duplicates is a trivial problem as a simple comparison between the two articles is sufficient. All whitespace is stripped from each article and passed through a hash function, after which a comparison between the two integers is carried out. If equal an exact duplicate has been identified and the later instance is excluded. These duplicates represent occasions when the data-source as simply reprinted an article word for word; the cause of such instances is unclear. On occasion such duplicates appear on the following day or may be the result of flaws within the repository archiving. The headlines of such text are ignored within the string comparison however as there are instances whereby the sole changes made within an article occur purely in the headline. These alterations are most likely due to an editorial decision regarding style rather than conveying any alteration of the given information. Within a sample corpus from the AP financial news wire retrieved from LexisNexis an instance of this occurring was identified whereby:

> *Rates rise in weekly Treasury auction. May 1st 2006*
>
> *Rates in weekly Treasury auction* ***hit highest levels in over five years.*** *May 1st 2006*

This alteration in the headline represents the sole difference between the two given news items; while the later does provide some form of new information regarding the extent of this rise, given the maintaining of the articles' content this appears to be perhaps a more sensationalist manner of reporting selected for journalist style alone.

Begins With/Ends With

These collisions represent occasions when a block of text has been inserted at the beginning or end of a previously released text; generally these items represent an expansion of some event

or greater level of explanation as demonstrated in figure 2.5. The provided excerpts of two news items within a retrieved corpus illustrate instances whereby a previously released news item is expanded upon; indeed the complete text of the second news item is approximately twice that of the original.

Bush **Trumpets New Economic Numbers-** AP Financial Newswire January 6^{th} 2006	Bush **Confident About Economy for 2006-** AP Financial Newswire January 6^{th} 2006
Bush, who also visited the Chicago Board of Trade, spoke as he and leaders of his economic team fanned out to trumpet recent improvements in the economy despite Friday's mixed jobs report showing a slowdown in monthly hiring. [...]	Bush, who also visited the Chicago Board of Trade, spoke as he and leaders of his economic team fanned out to trumpet recent improvements in the economy despite Friday's mixed jobs report showing a slowdown in monthly hiring. [...] **By highlighting recent economic advances, Bush took an opportunity to turn attention away from the conflict in Iraq. Bush and other members of his economic team emphasized the overall creation of**[...]

Figure 2.5: Begins with Ends with Collision.
An example for two collisions found within the news corpus. The text in bold was added to the original news item.

The identification of these events is achieved by comparing the two items where the longer item is trimmed; firstly from the end removing the text provided in bold, so that the two items are now of the same length and compared to the shorter such as in the above example this allows for the identify of a news item which has been expanded upon. A second comparison trims the beginning and allows for the identification of instances where some text is added to the beginning. If two given items are identified to begin or end with the same text the short is excluded as the longer expanded version should provide more information.

Longest Common Substring

While the two previously mentioned approaches do succeed in identifying some instances of duplicates neither is capable of identifying more complex duplicates. If a single character is altered within the middle of an article both methods would fail. This is problematic as these instances would be the most important as such duplicates are for a purely grammatical reason or journalistic style.

For the identification of these forms of duplicates a longest common substring approach was implemented. The longest common substring aims to locate the longest common sequence

within two or more given strings and is a common problem within the domain of computer science.

Stocks **Set to Open Down** on Oil Prices- AP Financial NewsWire January 12th 2006	Stocks **Open Lower** on Oil Prices- AP Financial NewsWire January 12th2006
The dollar declined against the Japanese yen, recently at 113.59 yen from 114.24 yen late Wednesday, ahead of U.S. trade data that's expected to show a narrowing deficit to $66.2 billion in November from $68.9 billion in October.	The dollar declined against the Japanese yen, recently at 113.59 yen from 114.24 yen late Wednesday, ahead of U.S. trade data that's expected to show a narrowing deficit to $66.2 billion in November from $68.9 billion in October.
Weekly jobless claims and import price index data are also due for release.	**The U.S. trade deficit improved slightly in November but was still the third highest on record as imports of foreign cars hit an all-time high and America's** [...]
The Euro and the British pound were [...]	The Euro and the British pound were [...]

Figure 2.6: Longest Common Substring Collisions.

The simplest implementation converts two given strings A and B into arrays; from this an operator may iterate through each element of the arrays and compare each element. The number of characters of the same value is recorded and the position within the given strings is stored. If the length of the longest common substring is greater than some threshold value based on the total length of the given texts it can be concluded that a duplicate has been identified. This method would be conservative in regards to the identification of duplicates; the length of the common substring must be greater than 60% of the shorter text before it can be concluded that such an event is a duplicate as shown in equation 2.7. This threshold value was set to a high level so that false negatives would be favoured over over false positives so as to reduce the risk of eliminating a large number of texts which were in fact not duplicates.

$$\frac{Length_{LongestCommonString}}{ArticleLength} = \begin{cases} > 60\% & \text{Collision Detected} \\ & \text{No Collision} \end{cases} \tag{2.7}$$

While this method is generally applied towards short strings it is possible to implement across longer strings such as two given news articles. This method for duplication detection also allows for the identification of instances where a small amount of text is inserted within the middle of a document being the only difference.

Examining the extract from the two articles presented in figure 2.6 a single paragraph has been changed within the centre of the news item however the rest of the text is identical. This gives and interesting example of a duplicate where the earlier article is of a speculative nature *Stocks Set to Open Down* while the second being a conformation *Stocks Open Lower*. The modification within the centre of the news item is an expansion, where new information

which was scheduled has been released and the original piece has been modified to include such information. When such collisions are detected if duplicates are to be excluded the first would be the ideal selection, given that the later would generally provide more information and would be of greatest interest, or if the exact timing is unknown the longer news item would be selected. As this method is influence by the total length of the article where a collisions is measured identified by the percentage of the text that is the same rather than the absolute length the length of the article should not play a large role in the identification of a collision. The biggest requirement is that there is a block of text larger than the threshold value that is uninterrupted. Where this method is most likely to fail is if there are number of small changes in the text. If however two articles were to include an extensive quote on the same topic making up the majority of the text but draw very different conclusions and commentaries provided the quote makes up a sufficient quantity of the text it would be flagged as a collision despite the different interpretations. If only a single character or value is changed at regular intervals in the text while the majority of the text might be identical the longest common sub-string will be interrupted and never represent more than a small percentage of uninterrupted text.

Levenshtein Distance

The Levenshtein distance algorithm is a string comparator which provides a metric for the difference between two string; the number of individual characters which need to be edited, inserted or deleted so as to make two strings identical developed by Vladimir Levenshtein [60]. A common application of this algorithm is within the area of spell-checking in text processing software and search engines to compensate for human errors in spelling and input [96].

A simple example such as the distance between Buffoon and Baboon outlines the methodology of the distance algorithm; the distance between the two is 3 as the replacement of 3 characters is required to make both strings equal.

1. BUFFOON $\rightarrow$ BUFOON - Removal of **F**
2. BUFOON $\rightarrow$ BU**B**OON- Alteration of **F** to **B**
3. BUBOON $\rightarrow$ B**A**BOON- Alteration of **U** to **A**

Collisions detected through the Levenshtein distance algorithm are those generally arising from numerous minor modifications throughout the document, as seen in Figure 2.7 there is minimal difference in regards to the actual content of the pieces. If a number of minimal differences are constantly encountered the longest common substring would be rather small in length, and would fail to be identified by the proposed methodology. Differences stem from updated values regarding market prices; however the major text within the document remains the same. While the later does provide updated and relevant information, the textual

information has no difference, as our interest is in textual and qualitative analysis these figures would have no impact on any analysis but the duplication of textual information would lead to double counting between the two articles.

Also of note is the altering of the headlines across the articles, while the first refers to the increasing in price due to positive namely growth, the later shifts the focus to a fearful headline over the risk of international conflict, both however maintain the same discussion regarding Iran whereby the key tone and discussion of the articles are unaltered other than an update regarding current market prices.

Oil Prices **Up on Expectation of Growth**- AP Financial NewsWire January 12^{th} 2006	Oil Prices **Climb on Fears Over Iran**- AP Financial NewsWire January 12^{th} 2006
Oil prices rose Thursday amid market jitters over Iran's nuclear development and on traders' convictions that economic growth will cause energy consumption to rise. Light sweet crude for February delivery rose **96** cents to **$64.90** a barrel **in** **morning trade** on the New York Mercantile Exchange. Heating oil futures jumped **2.2** cents to **$1.7493** a gallon, while gasoline surged 2 1/2 cents to **$1.7576** a gallon. Natural gas **futures rose 4.2 cents at $9.254** per 1,000 cubic feet. The threat of instability in the Middle East [...]	Oil prices rose Thursday amid market jitters over Iran's nuclear development and on traders' convictions that economic growth will cause energy consumption to rise. Light, sweet crude for February delivery rose **72** cents to **$64.66** a barrel **by** **midday in Europe in electronic trading** on the New York Mercantile Exchange. Heating oil jumped **over 2** cents to **$1.7480** a gallon, while gasoline surged 2 1/2 cents to **$1.7579** a gallon. Natural gas **gained 6 cents to $9.300** per 1,000 cubic feet. The threat of instability in the Middle East [...]

Figure 2.7: Levenshtein Distance Collisions.

Clearly the distance between the above articles is minimal as when compared to the total length of the articles only a few numerical characters must be altered, similarly to the longest common substring, a threshold value must be determined. This value would similarly be based on the total number of alterations when compared to the overall length of the news items. If the number of edits required is less than 10% of all characters within the longest document it is concluded that a collision has been detected as shown in equation 2.8.

$$\frac{Distance}{ArticleLength} = \begin{cases} < 10\% & \text{Collision Detected} \\ & \text{No Collision} \end{cases} \tag{2.8}$$

2.4.3 Discussion

This section has discussed the instances of collisions; the reprinting of news items verbatim or the expanding and modifying of previously released news items. While such re-releases may in themselves contain no new information the act of repetition and expanded discussion on a single event may indicate the importance of such an event or topic. The inclusion of any such instances may have a significant impact on any textual analysis, if the reprinting and expansion of a news item indicated importance the inclusion of an event such news items would be of key interest. By evaluating the changes between individual collision it is also possible to view the changes in the reporting of an event over time. Articles may simply add information but as seen may also shift the tone of the news piece and the intended direction and bearing of the commentary.

2.5 Evolutionary Computing

The following section presents an examination of two forms of evolutionary algorithms; examining their origins and later expansion into numerous applications across various fields of computer science and information science while examining their practicality within textual analysis and lexicon development. Two evolutionary algorithms are applied in the interest of completeness and to test the methodology with two algorithms which have both been shown to perform well for feature selection.

Biologically inspired algorithms is an area of algorithms which take inspiration from behavior seen within areas of nature. The field is composed of a number of algorithms which draw inspiration for real world observations from social, biological or genetic behaviour; such works have frequently originated from a desire towards the modeling of such events with the aim to gain a greater level of understanding of the mechanics behind such events. Yet subsequent research has displayed their success across a number of computational problems.

Firstly is a discussion on genetic algorithms, developed by John Holland [39] which draw inspiration from evolutionary genetic behaviour witnessed within nature. Beginning is an overview of their origins and key concepts for consideration, examining the mechanisms by which genetic algorithms mimic life and the process in which the system learns. Following is an expansion of the mechanics of the system and an examination of their practicality and implementation within the field of feature selection and a number of varying situations. Secondly is examination of particle swarm optimization, initially developed by Kennedy & Eberhart; particle swarms represent an alternative area within evolutionary computing which draw inspiration from social and group behaviours. Groups of animals such as flocks of birds and schools of fish are witness to behave in some form of complex group behaviour, while

initially aimed at simulating real life behaviour such algorithms have since proved robust within areas of optimization and problem solving.

My aim here is to apply genetic algorithms and particle swarm optimization to the area of lexicon refinement for sentiment analysis of financial news. By using a reference lexicon like the General Inquirer my aim is to identify terms which in finance do not have a positive or negative affect meaning. By giving the system a list of what have been defined as affect bearing terms the aim is to select key terms and return a word list which is a subset of the initial list, this new refined word list should provide for a better measurement of sentiment in financial news.

Genetic Algorithms were selected as they have been shown to be a highly effect problem solving techniques, and perform well in the area of feature selection. In particular they perform well in areas with a high number of features and noisy data [112], including the area of feature selection in textual analysis. Genetic Algorithms have been successfully applied to the areas of text mining [6], topic identification [48] and text classification [82].

Abbasi et al. [1] used genetic algorithms for feature selection of textual analysis of movie reviews and U.S and Middle Eastern web forum postings. In their work the solution produced from GAs was a list of terms used for text classification. An initial feature set was extracted from a corpus of files, the solution produced by the system was a subset of terms, which by using they improved the classification accuracy of the system. This method was similarly used by Mukherjee et al. [72] for content analysis of emails. Documents were summarised as a vector of terms and their frequency, these were compared against GA solutions where terms were excluded if the feature representing that term in the solution was set to 0. In both cases the produced solution will represent a list of relevant terms for the given task.

A deeper discussion regarding the implementation of genetic algorithms and particle swarm optimization within the context of textual analysis is presented in section 3.3.

2.5.1 Genetic Algorithms

Genetic algorithms are a search heuristic and optimization methodology which draw inspiration from natural selection and genetics, implementing genetic operators witnessed within nature, such as the concept of survival of the fitness, breeding and mutation.

Genetic algorithms begin with a population of candidates, each of which is a proposed solution to a given problem; each candidate represents a solution as a binary string of size n. Upon initialisation the encoding of each candidate within the population will be generated in a random manner with each bit within the candidate being randomly set to 1 or 0.

Upon each iteration or generation, each candidate solution is evaluated against a given fitness function; from which a *fitness value* may be assigned determined by the solutions

performance at solving a given problem. Once evaluated a limited number of candidates are selected for survival; the probability of which is determined by the relative fitness value of each candidate. Through this competition the system ensures that candidates which have performed well will continue to survive and be selected as parents. Selected parents will be paired and components of each will contribute to the breeding of a child candidate for the following generation. Candidates well adapted to the given environment(problem) will, as seen within nature, survive and pass on their genetic material/information towards the generation of new individuals.

Successful traits and characteristics will spread throughout the population while those which perform poorly will fail to be passed on, such robustness provides an attractive methodology towards the adaptive development of solutions for complex optimization and search problems within computing. Successful individuals will pass on their genetic information and combine with those who have also prove successful, from here valuable information and complex solutions will overtime emerge.

Early work within genetic algorithms was generally concerned with the development of algorithms allowing for a simulation of real world genetic behaviour, hoping to give greater understanding towards the role played by genetics in the development of organisms capable of adapting to a given environment.Later work by Holland [39] explored the implications of genetic algorithms and their ability to *evolve* solutions towards complex problems. Genetic algorithms proved as a means of providing a robust approach towards problem solving, whereby through their adaptive nature complex and successful solutions would emerge during the course of the programs execution, while being relatively simplistic to implement they have proven to be highly adaptive and robust. Genetic algorithms have been shown to perform well across a number of real world problems across various domains including; artificial intelligence, computer vision, timetabling, stock market analysis [79,106], engineering problems [29,74,95] , medical diagnosis [112] and learning mechanisms for video game AI [15,65].

Genetic Operators

An overview of the stages involved in implementing a genetic algorithm are outlined in figure 2.8, to begin with a population of size N candidates are produced, the initial generation is randomly generated with each bit within candidates being set randomly to 0 or 1. Following, all candidates will be evaluated in regards to their performance towards the solution of a given problem, from which each may be assigned a *fitness value*, this value may be considered analogous to an individual's performance in nature and its ability to adapt to a proposed environment. At this point there is now some measurement regarding the relative level of success of each candidate, after the process of generating a new population begins.

Through some selection method a number of individuals will be selected who will contribute

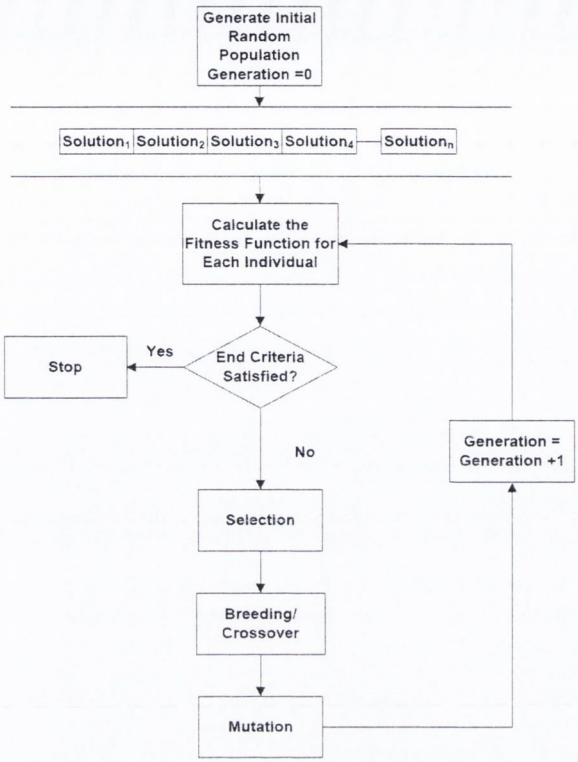


Figure 2.8: Genetic Algorithm Flow Diagram.

towards the generation of a new population, the probability of an candidates survival and selection is dependent on their relative fitness when compared to the population as a whole. Once a sufficient number of individuals have been selected they will be randomly paired for the purpose of recombination whereby according to a *crossover scheme* two candidates will be combined, drawing elements from each selected parent producing one or more *offspring*. Subsequently and in further mimicry of natural genetics, a very small number of candidates will be selected for *mutation*, this change will modify a candidate in a small and completely random manner, as in nature the frequency of such occurrences is very rare and allow for the introduction of variation within a given population and may be beneficial or detrimental. This process will continue, with each iteration (generation) producing a population of new candidates from which superior candidates will emerge, flourish and which should produce some form of optimal solution. This process will continue until some predetermined end criteria such as number of generations, convergence of the system or performance level has been met.

While the main components of genetic algorithms initially appear simple; based on weighted random selection & merging of strings between two individuals, yet over time complex and successful solutions emerge, producing superior offspring and provide a valuable method towards the solution of complex problems. While the initial generation is a purely ran-

dom search mechanism, subsequent generations incorporate elements of randomness towards a *"highly exploitative search through a coding parameter space"* [34].

Examining the problem space in figure 2.9 a number of peaks are presented however poor configuration may cause the population to converge towards one of the sub-optimal peaks, if the population begins to converge exploration of further problem space may not occur resulting in a poor solution when compared to those available. Ideally early populations would cover a wide area of the problem space, an evaluation of each point with the problem space would be infeasible however if a population is wildly distributed across the space a number of candidates will group towards each of the peaks however with adequate control in regards to convergence the early discovery of a peak would not begin to dominate the system. As the population ages the *true* highest peak (if one does exist) will be explored in greater depth.

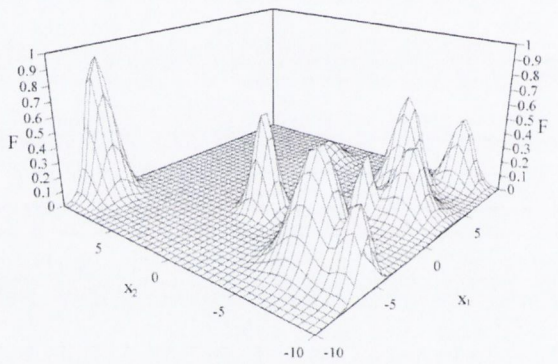


Figure 2.9: An Example of a Multi Peak Function [26].

A potential problem which may arise through poor configuration of the population is that if selection intensity is too high the population may begin to converge too early on a sub-optimal peak, the problem space in figure 2.9 contains a number of such sub-optimal peaks. A notable area of research is the development of methods so as to mitigate the chances of such events occurring. Generally these issues may be countered through a dynamic configuration of the system which will encourage a greater deal of exploration and delay the risk of early convergence until later developments, thus allowing for solutions which may perform poorly during early generations to survive thereby expanding the diversity of the system [71].

An alternative proposed method is for the allowing of new randomly generated solutions to enter the population [25]. These *random immigrants* [14] would be analogous to the arrival of a new species or race within a real world genetic environment, such arrivals may occur periodically throughout the system's lifetime or if the diversity of the population drops beneath a predetermined level before a sufficient number of generations have passed. Finally GAs rarely have a means for considering the long term effect of minor changes, such as the sacrifice of performance for a single generation which however leads to an expansion of system diversity

leading to superior performance as a whole. While also being unable to correct for detrimental effects on solutions through breeding as a GA system rarely implements a memory whereby the effect of each solution based on its breeding is tracked. A method for countering such an issue is through the inclusion of *elitism*, whereby the solution with the highest performance passes from one generation to the next unchanged by mutation or crossover. Such implementation is found to be beneficial to the system's performance.

Feature Selection & Lexicon Generation Through GAs

While much work has be carried out within the field of feature selection/reduction [73, 77, 113] within textual analysis much of this work relies on expert knowledge of systems for the reduction of terms within the feature set based on the users expert knowledge. However where implemented GAs have been shown to be effective towards feature selection/reduction within textual analysis and sentiment analysis [1]. Genetic algorithms tend to perform well in systems with a small number of variables and search space with a limited number of variables, however as the complexity of the system grows so does computational expense of the system while also being detrimental to the system's performance. Furthermore the increased encoding size leads to issues in terms of breeding and mutation [63, 71, 114, 115]. The configuration of the system itself is a non-trivial problem [35, 99], where an appropriate selection on population size, mutation rate and crossover rate play a key role in the performance of GAs, and while a number of studies have been conducted towards the examination of appropriate probability values, there is no generalised method for variable selection for the optimal system configuration.

Considering an optimization problem whereby there are a number of inputs n and the goal is to engineer a solution to maximise the function $g = f(x)$ where x is the configuration of inputs which are set to true. As each input may be set to true or false, this will results in a total of 2^n possible configurations. Where n is small a brute force attempt would be simple however as n increases the solution space grows exponentially. This problem may be easily encoded into a binary string and evaluated using a GA. This begins by generating x candidate solutions with binary strings of size n. If $n = 6$ an example of an initial randomly generated solutions could be considered as:

$$A_1 = 011110$$
$$A_2 = 110010$$

Whereby a 1 indicates the respective input is set to true and 0 indicates it is set to false. When this is applied to a bag of words technique in text analysis GAs can be used to find terms of interest. Beginning with a bag of words approach where the frequency of a set of terms from

a lexicon are recorded each individual term can be viewed as a feature. GAs can be used to remove terms which in a given classification process are irrelevant. When considering the collapsing of a document into a vector when using a bag of words approach it is clear how this method of representing solutions is well suited to feature selection in textual analysis. Once a bag of words vector has been composed it is possible to implement the previously mentioned steps for the evolution of a candidate solution which results in a vector of relevant features as defined against the fitness function regarding desire to summarise and extract meaningful data from the given corpus. These candidates would each represent a list of terms which is a subset of the terms in the reference lexicon. Each solution would give a list of terms which are relevant for the text classification and, remove terms whos frequency would not play a role in the text classification problem. Once the population has converged a new word list made up of a subset of the term found in the initial lexicon would be produced which should result in improved text classification. When applied to sentiment analysis, given an initial list of positive or negative affect term candidate solutions would aim to remove terms which have either no affect meaning in the target domain when compared to general language. A term may have a positive or negative meaning in the general language but in a certain domain would either have no affect bearing or could be entirely misclassified.

The largest challenge with such an approach to textual analysis is the issue of a defined fitness function. So as to implement genetic algorithms across a given problem a method for measuring the effectiveness of the proposed solution is needed. Within text categorization this may be the accuracy of the resultant output however as with all such training methods a large pre-labeled corpus is required, such work is (generally) carried out through manual evaluation and as such is time consuming and open to disagreements between participants.

One issue regarding the application of such a method towards document analysis is that each unique term within the document will be another feature within the solution. While this posses some challenges, studies [71] have demonstrated the robustness of genetic algorithms towards solutions in large scale feature space with modifications towards the breeding process so as to encourage adequate feature space exploration.

Within the field of feature selection an issue to consider is the occurrence of features not present within the training set but may be found within the blind testing data. If a feature is not present within the training set its inclusion in any solution will have no impact on the candidate's performance yet may be detrimental to performance once this solution is deployed against blind data. A proposed means of countering this through the implementation of a *Penalty Function*. whereby the complexity of the candidate has an impact on the final fitness value. Consider two candidate which have the same initial fitness value, however within candidate A feature n is found to be true while in candidate B feature n is found to be false it can see that this feature has little to no bearing on the candidate's performance yet as it may be

caused as a result of being an absentee feature within the training set this may prove detrimental once examined against blind data. In the given example a scaled fitness value for B would now be higher than A, thus the inclusion of feature n within candidates would be discouraged. However the implementation of penalty functions can be problematic small changes towards the penalty coefficient may result in unpredictable, and decremental consequences [41]. A penalty function may also be implemented so as to account for constraints which may be found in an optimization problem, while a penalty function may provide an effective manner for dealing with solutions which violate any constraints issues remain. Firstly the implementation of an appropriate penalty function may not always be simple, and secondly computational time may be wasted in instances whereby invalid solutions are still evaluated [66, 116].

An alternative implementation of a penalty function may be directed towards the encouragement to explore a greater range within the search space. Solutions which concentrate on areas previously *visited* by other solutions would be penalized. Such a penalty would decrease the fitness value of the given solution encouraging the expansion into unexplored search space areas [70]. Such fitness sharing [24, 34] treats fitness as a shared resource, therefore as the number of solutions within close proximity increases the solutions fitness is worsened encouraging solutions to spread out and explore further regions of the problem space. However such a method posses the risk that while at some point it is desirable that convergence occurs so as to concentrate exploration within a smaller region where the greatest performance has been observed. This issue may be countered through the usage of a dynamic aspect to the penalty function based on time, whereby as the population ages the impact of this penalty decreases. Such concepts will be examined in depth in section 3.3.

Information Exchange and Destruction

Goldberg [34] proposed that if each candidate is a string of size n encoded according to some alphabet each complete string may be considered as an idea, while sub-strings may be considered as notions. Survival may be considered as the propagation of an idea, recombination as the exchange of notions and the emergence of new ideas, as with individuals this could be seen as a form of communication, and the propagation of an idea as agreement. If the population begins to converge such that all members are minor variations of a single candidate a consensus has been met, while the implementation of mutation is the exploration of new concepts, ideas or experimentation. Superior candidates are those who's notions and ideas best describe the problem space or adaptation towards the proposed environment. A key area of importance within GAs is determining a balance between *Exploration* and *Exploitation*, candidates will need to be given the ability to develop and propagate, while ensuring that the population does not become stagnant with little exploration of problem space area.

During early generations it may be desirable to allow for the survival of individuals who

may otherwise die, simply so as to encourage the exploration of a greater area of the problem space. If a small number of candidates emerge early with a relatively high fitness value, they will be selected far more often for breading and allow for early convergence of the population whereby little diversity exists. Therefore it may be desirable to encourage the population at an early stage to explore the problem space in greater depth. Conversely while it is desirable to explore wide areas early on at some stage it is beneficial that the system converges towards a smaller area where the best performance has been observed and explores the area in greater detail. The interaction between two *parent* candidates has the potential to produce a new solution which while composed of both parents bare little resemblance to either [20, 32]. This represents a high level of information exchange and poses the risk of resulting in the *destruction* of previously developed notions and ideas. Therefore at times it may be beneficial to allow for a greater exchange from one parent over the other the implementation of which and implications will be presented later in section 3.3.1.

2.5.2 Particle Swarm Optimisation

Following is an examination and contrast of genetic algorithms with particle swarm optimisation(PSO); an alternative optimization technique within the field of Evolutionary Computing. While genetic algorithms draw inspiration from the behaviour of genes seen in nature, the roots of particle swarm optimization stem from a mimicry of swarm behaviour seen within nature such as the flocking of fish and schools of birds, where a collective of individuals behave in a group and co-operative manner through some for social interaction and learning.

Overview

Particle swarm optimization is a search optimisation technique developed by Kennedy and Eberhart [23,45,46] which draws inspiration from social behaviours seen within nature, namely flocking of birds and schools of fish and form another area of biologically inspired computing. Similarly to genetic algorithms the precursor to particle swarm optimization was aimed at the modeling of real-life behaviour; in this instance the flocking and group movement of animals primarily birds flocks or schools of fish. Researchers had noted that through some unknown *Rules* these groups may travel in a carefully choreographed movement; moving as a single entity, changing direction and velocity yet when required scattering and later regrouping.

Graphical research aimed to develop a means for the simulation of large flocks of animals but noted that traditional methods of scripting a large population would be tedious and open to errors. By modeling a flock as the sum of individual behaviours of entities the researchers aimed to develop a system whereby the behaviour of a flock would be determined by the impact of each individual *boid* (Bird like object) on the population as a whole [84]. While this work

did not argue to provide a true representation of flocking behaviour the researchers observed that through the inclusion of a number of simple rules:

- Collision Avoidance: Avoid Collision with nearby entities
- Velocity Matching: Match Velocity with surrounding entities
- Flock Centering: Attempt to stay close to nearby entities

complex behaviour emerges.

This is an example of emergent behaviour whereby over time a system with a simple number of rules will overtime produce complex and at time unpredictable behaviours. The authors went as far as to argue that due to the "detached nature of control" the creator becomes less of an animator but rather a *meta-animator* observing that "these darn boids seem to have a mind of their own". This is a common outcome within evolutionary computing where similarly to GAs given the guided yet random nature of the algorithms implementation and progression overtime, whereby without prior knowledge complex behaviours and solutions can emerge over time. These methods for the simulation of artificial groups was further expanded upon, frequently in the area of simulating schools of fish [43, 109], however such works are graphical or ecological [86] in nature and tended to concentrated purely on the improved nature of graphical modeling such as the inclusion of various levels of physics and real life situations and further criteria rather than investigating the potential of implementing such techniques towards machine learning.

Kennedy and Eberhart expanded upon these techniques to examine the practicality of implementing such a simulated social model towards machine learning for the purpose of neural network training. The proposed PSO algorithm begins with a given number of *Particles*; which similarly to GAs are spread randomly across a problem space upon initialisation, each individual particle is represented by a vector X of size D of real values indicating its current location within the problem space. Secondly each particle contains a velocity directing its future movements within space, upon each iteration the location is updated and the particle's performance or *fitness* is evaluated after which the particle's velocity V_{id} and position X_id is updated according to equations 2.9 & 2.10 across each dimension. Particles also contain a *memory* of the location in which its most successful performance to date was observed P_{id}, the recall of this location is considered to represent an individual's personal cognitive learning through experience [23], while a secondary location P_{gd} being the location where the best performance was observed across the swarm as a whole. This communal exchange of information may be considered a social learning mechanism, whereby information is exchanged across the population as a whole.

$X_{(i,d)}$	0.2	0.9	0.6	0.3	0.4	0.7	0.3	0.6	0.8	0.1

Figure 2.10: An Example of Particle Swarm Encoding

$$V_{id} = V_{id} + \overbrace{C_1 * \varphi_1 * (P_{id} - X_{id})}^{Cognitive} + \overbrace{C_2 * \varphi_2 * (P_{gd} - X_{id})}^{Social} \tag{2.9}$$

$$X_{id} = X_{id} + V_{id} \tag{2.10}$$

C_1 & C_2 are two constants where *generally* $C_1 = C_2 = 2$, similarly to GAs a random number φ allows for an aspect of randomness towards problem space exploration yet is guided so as to examine promising locations within the problem space.

Figure 2.11 illustrates that the movement of a particle from one iteration to the next is a balance between the particle's previous velocity, its own experience and previous learning with P_{id} and the influence of the swarm as a whole through P_{gd}. As opposed to genetic algorithms, within particle swarms individual entities are not replaced but rather are updated upon each iteration, the inclusion of the cognitive learning allows for the particle to *remember* where it has been, while the inclusion of its velocity from the previous iteration allows that its behaviour will not be changed drastically from one iteration to the next. Such personal information aids in the prevention the large knowledge destruction which is possible within genetic algorithms through crossover. While particle swarms are based on modeling the physical location of an object in space these particles are simply a vector of real values. So while developed for the purpose of modeling the behaviour of physical object when applied to problem solving the methodology is simply a techniques in machine learning for calculating weights or inputs. This is shown in figure 2.10 with the values for the location of particle d across a ten dimension problem space with values bounded between 0 and 1.

While particle swarm optimization has been the focus of numerous studies, much of the framework put forward by Kennedy and Eberhart remains at the heart of their implementation. Much of this research has been devoted to a further understanding of the exact nature by which particle swarms learn and change their behaviour over time [13]. Similarly to GAs particle swarms must ideally find a balance between exploration and exploitation [108]. Later works aimed towards developing means so as to vary the levels of cognitive and social learning's impact on the particle's behaviour, similarly to GAs a risk is that the cognitive learning will over time influence the early convergence of swarms towards a sub-optimal point within the problem space [93].

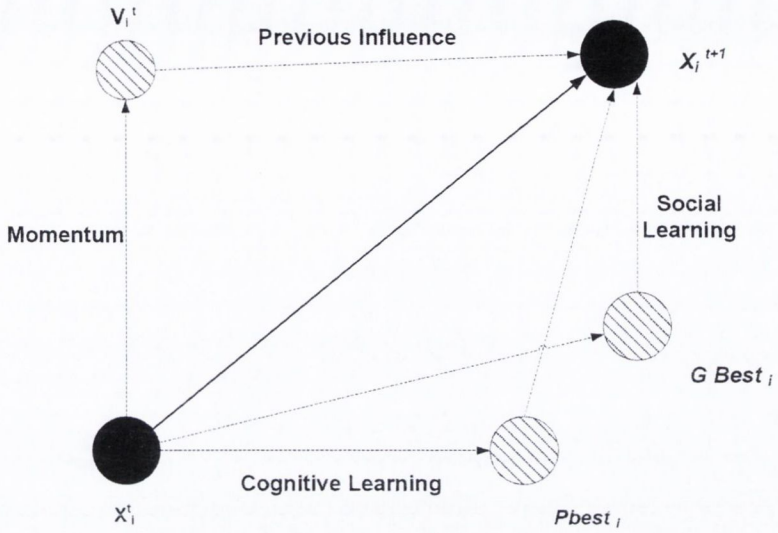


Figure 2.11: Impact of Social and Cognitive learning on a Particle's velocity.

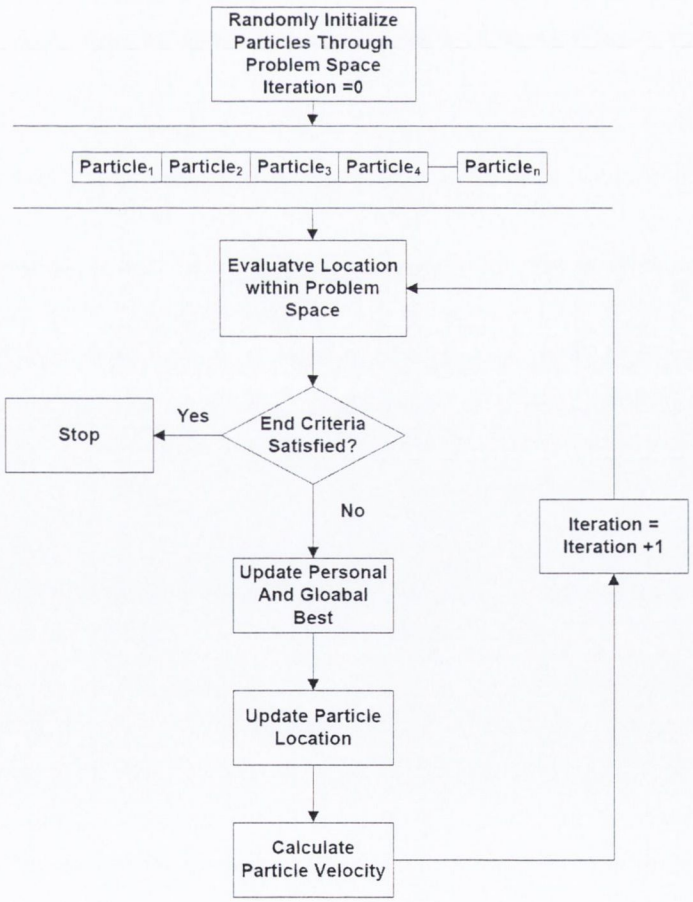


Figure 2.12: Particle Swarm Optimization Flow Diagram.

Figure 2.12 illustrates the implementation of a particle swarm and demonstrates the simple steps involved which through the usage of the above equations yet produce robust solutions. Here the three stage process of improvement within genetic algorithms; selection, breeding and mutation have been replaced with a recall of optimal positions and an update of individual particle's current velocity.

Concept Learning

A further examination of particle swarms and their behaviour as a simulation of the exchange and refinement of knowledge within human society [45], noted that the method by which a particle *learns* to move through the problem space contains parallels in regards to our own manner of absorbing and assimilating information. Consider the manner in which people learn information; a social communicative aspect is required, people must engage with other individuals and exchange information towards the pursuit of learning be it through "literature, pedantry, informal conversation" each requires that an individual engage in some form of a communicative role. Yet it is from this *communal* exchange of information, methods and notions that personal ideas and notions begin to emerge, our social information exchange is refined through personal experience and information processing.

Through repeated simulations whereby a particle learns purely on a individualistic or social manner were examined and compared. The role of social learning was identified to be of key importance within the system's performance, a selfish cognitive only model suffers from the fact that as a particle is encouraged to remain towards its own past location the ability for wide exploration is hampered as there is a continuous draw or encouragement for the particle to concentrate purely on where it has been and what it has experienced, indeed a cognitive only based system is identical to iterating a swarm of size 1 repeatedly.

For an evaluation of the social impact of learning a modification may be made such that rather than a global exchange of information whereby there is a single $gbest_i$ a local or "neighbourhood" value can be considered [23]. A simplistic manner for defining what constitutes a "neighbour" is to simply evaluate those particles which are found in proximity within the array or collection structure. While considering those particles found adjacent to a given particle with an array or list as truly "neighbours" as each has been randomly spread regardless of the chronological order of their initialisation, over time as all are exposed to a similar pull towards a shared *local best* $lbest_i$ a rather truer neighbourhood emerges. These individual neighbourhoods can be considered as cliques within the swarm, exchanging information and communicating within a small sub-group of individuals. Alternatively the distance between each particle could be evaluated in a simplistic manner, however such a method may add to the computational time while also running the risk that the "neighbours" change frequently, such a method would negate the benefits of which this sub-populational

learning is aimed.

The velocity component whereby the particles movements are influenced further by those directly proceeding allow for the encouragement of constant movement, while being drawn towards the regions where a high level of performance has been observed. As velocity is a component defined influence through each previous iteration it is a continuous learning component of the particle, being the product of all previous learning experienced, both cognitive and social.

Feature Selection

Similarly to GAs if a documents is collapsed into a vector V representing the frequency of terms, particle swarm optimization may be suited for the task of feature selection within textual analysis. The original implementation of PSO allows for the production of a vector V across n dimensions of real values, if this is applied to the problem area of feature selection such a method allows for the evolution of weights for each individual input. If the aim is to apply such a method towards textual analysis these weights provide information regarding the relative importance or impact of the term in question [62]. The solution produced through this methodology was a vector of value bounded between 0 and 1 which were applied as weights, as shown in figure 2.10. This is in contrast to genetic algorithms which as outlines in section 2.5.1 the produced solution is binary in nature, indicating the inclusion or exclusion of individual terms, yet through this method alone will not produce an individual weight towards each term. However further work on PSOs have examined the refinement towards a discrete binary version [47], the initial framework for calculating the velocity and position or a particle remains unchanged, however all positional values P_{id}, P_{gd} and X_{id} are now discrete integers in $[0, 1]$, this is generally achieved through a probability function applied to map the continuous value to a binary space.

As with GAs an area of concern is that while particle swarms have been shown effective within a number of optimization problems early work concentrated on relativity small feature spaces with a limited number of dimensions. However researchers have examined the ability of particles to perform within high dimensionality problem spaces, and methods through which to improve this [38, 61].

2.5.3 Discussion

Two forms of biologically inspired algorithms which have demonstrated to be highly successful within many areas of machine learning and optimization have been outlined. This work aims to apply these methods to feature selection of legacy lexica. If a document is summarised as a term frequency vector of size n, by using either GAs or PSO through machine learning

techniques it may be possible to identify terms which are of importance for textual analysis. Both GAs and PSO represent solutions as a vector of values and while particles tend to store numeric values these can be mapped to a binary space. While there are issues regarding the implementation of these methods both have been shown to be highly robust within optimization problems across a wide range of fields. While both suffer from issues are the number of variables/features grow, much research has been concentrated on implementing them across a large feature space with successful results. Their implementation towards sentiment analysis, specifically directed towards financial news will be discussed in section 3.3.

2.6 Summary

This chapter has provided a brief outline regarding the field of textual analysis, discussing a number of its application and resources available. The major theme has been a discussion on the importance on the domain to which this analysis is aimed. While large scale complex lexica are advantageous and powerful tools, their success across all domains is not guaranteed without some form of intervention or refinement. While those lexica produced through expert human development towards a specific domain offer an alternative source of information, as noted such lexica are not only difficult to produce for each individual domain, but are open to flaws. If sentiment is contained within a text it is encoded through a selection of key words by the author, identifying, these allows for for abstraction of such sentiment, quantify it and evaluate it as a proxy, based on the frequency of these key terms. These key terms would be the result of the authors understand of the topic domain. An examination on the corpus itself must be carried out, allowing for some form of extraction and machine learning through the usage of pre-annotated documents.

From this the flow of news and information over time is examined, proposing a methodological approach for the identification of collisions within a corpus of news items. Proposing an approach towards the identification of duplicate news items which occur over time, noting that while some changes may be trivial others represent significant alterations to the tone and intended message regarding the news item. Following, a brief discussion regarding the area of sentiment analysis and finance was proposed, where-in researchers have begun to examine the role which sentiment impact future market behaviour. Where sentiment within news itself impacts human perception of future market returns or merely captures such information in a linguistical qualitative manner rather than technical statistical analysis of market and company fundamentals is unknown. However as stated markets are not rational in their behaviour, and neither are human individuals with regards to their absorption of information and the impact of which sentiment has towards them. This is the field towards which the feasibility of automated lexicon refinement and its applications within sentiment analysis is examined, the goal of which

being to examine and potentially improve analysis which examine the explanatory power of an affect based time series on future market returns while also providing a refined lexicon which should in theory provide insight into the language and emotionally laden terminology within financial news reporting. The goal of this work is to achieve this through the implementation of the aforementioned biologically inspired algorithms, both of which by drawing inspiration from behaviour wittiness within nature have proposed powerful and adaptive means towards optimization problems; which here will be the refinement of legacy lexicon.

Chapter 3

Methodology

Stone et al. and the pioneers Lasswell, Namenwirth and others, have claimed, and perhaps demonstrated, that affect, and by implication sentiment, is expressed in words, and is then incorporated in the discourse of a specialist community; politics and lately social psychology then Lasswell and Stone areas of study respectively. The articulation of affect in text can be observed and some writers, especially fiction writers do so well. Primarily through speech, gestures, and facial expressions are thought to be the principal channels of affect and the kindred emotions.

Early studies in political science focused on the understanding of affect in the discourse of opposing political parties (US Republican and Democratic parties) and in the editorial opinions in the major national newspapers of countries in the Inter war period (1920-1940), namely the USA, UK, USSR and Germany. Political parties strive for power as do national governments. One might argue that in this contest, either of political parties or nation states, there will be a polarity of views expressed by certain parties.

Academic intuition led Lasswell, and then Stone, to suggest that a study of representative texts of a specialist domain, will help in the identification of polarity of opinions will help in the identification of the polarity of opinions amongst the annotators. If the text samples are small enough and the scrutineers capable enough, then the analysis is straight forward in the sense that the capable scrutineers will be able to reasonably extract the affect expressed in the texts.

What the early pioneers extracted suggested that computers be used to capture the "*affect effect*" as it is now termed, given that computers cannot read or interpret written texts per se, it was suggested that computers should be given a list of key affect words, and then computers could count occurrences of the affect words within the representative sample of text. This frequency of occurrences could be aggregated over the sample, and inferences be drawn about the dominance of a set of affect words. Lasswell had introduced intuitively appealing categories- negative and positive affect, power affiliation and so on (28 categories to be precise)

to which sets of words can be assigned. This assignment was non-exclusive generally in that the affect words can belong to one or more categories. Lasswell also included keywords from politics and science linked to politics like economics. There were at least three versions of the Lasswell dictionary of affect. This dictionary was subsumed by Stone; General Inquirer Dictionary. Stone had created a program, the GI, which could count affect words and deal with some of the ambiguity inherent in nature.

Stone notes that the extended GI dictionary- with 82 categories and 11,000 words, was created by the sample of experts proposing lists of affect words, and discussion then ensued which leads to the selection and inclusion of the words. Some of the words were found to be rarely used while others were more frequently used. One "automatic" selection criteria was to check how frequency this words was used in representative sample of texts, written in American English for the consumption of lay American English comprehending audience. Stone and colleagues used the frequency statistics used by the compiled of Thorndike (American) English Dictionary.

Over 40 years since the GI dictionary was compiled, Tetlock [104] used the dictionary almost in total to analyse opinion columns in the financial daily WallStreet Journal, Tetlock's purpose however, was different to that of Lasswell or Stone on two counts. First, Tetlock analysed 21st century American English texts that was primarily financial in orientation. Second, Tetlock used complex statistical regression of a few words in the negative sentiment category, as an exogenous variable impacting on the return on investment on the Dow Jones Industrial Average.

In more recent works on the analysis of returns on commodity markets, especially crude oil, Ahmad et al [3, 18] it was noted that some of the affect terms in the GI actually are not used for expressing sentiment polarity, attitude or orientation about crude oil in particular and financial markets in general. Terms like *"crude", "sweet", "light", "heavy", "active concentration"* in oil terminology and adjective used for qualifying objects and not for expressing affect. The same is true about key financial terms like *"company" or "shares"* which have no affect connotations. (Tetlock may have avoided this by using disambiguation in the GI system.) Ahmad's solution was to use GI, as it is psychologically and politically very well grounded, and in addition they used a specialist lexicon- with its subject specific categories- and the words common to the specialist lexicon and the GI dictionary, are not used in the computing of affect. In the last few years, some papers have criticised Tetlock's use of GI words, primarily on the grounds that some of the words are archaic and some are domain terms. It appears that the business of compiling affect dictionaries is muti-phase project: First, initiation, where a set of words are chosen on lexical, psychological and domain specific basis. Second these words are used in the analysis of domain problems. Third, the impact of the affect words, or their suitability of these words in describing people's or market's sentiment is determined by

using the frequency of these words as exogenous variables in the endogenous time series of (returns on) prices. Fourth, there is *pruning* process whereby some words are excluded and yet others are included. A method has been developed whereby the initiation, selection, fitness by ways of predicting the behaviour of endogenous variables, and pruning is conducted as a search/optimization problem.

Following is an outline of the development of a methodology for the refinement of legacy lexica, outlining the steps and processed involved in the collection, sanitisation, quantification and analysis of sentiment when applied towards financial news and indices. Firstly the development of a method allowing for the large scale collection and archiving of textual data is discussed; an examination is presented regarding the frequency of duplicate news items retrieved from a selected online repository. An outline follows of a method which allows for the generation of textual based time series from a given lexicon. This time series allows for an examination of the frequency of affect terms as provided from a reference lexicon. Closing is a discussion on the implementation of genetic algorithms and particle swarm optimisation to lexicon reduction. An explanation on the implementation and training of these algorithms in this approach to sentiment analysis. Using these two machine learning techniques the goal is to, in an automated manner, produce a reduced lexicon with is specifically tailored for the domain of sentiment analysis of financial news. This work applies these refined lexica and their development towards sentiment analysis of financial news, through an examination of the affect time series generated through a reference lexicon and market movements, aiming to use the market itself as a barometer for market sentiment to produce a refined lexicon of key affect terms.

3.1 Data Curation

Within this work the term data curation is used for the large scale collection, sanitizing and archiving of textual data in a consistent and automated manner. In any form of textual analysis a means for the rapid collection of information is key, and ideally minimal human interaction is required. With the expanding nature of the internet and large scale archiving of news and items there exists numerous online repositories which provide the ability to curate data from across a number of news sources. Content may be filtered according to source, topics, dates and nature language content, offering a means for the collection of large quantities of textual data for the purpose of corpus development. This section will focus on news flow, especially when news stories are repeated in part or whole. Some researchers use the total number of stories about a topic as a sentiment proxy. Since this work examines the frequency of terms, if news item are repeated by accident this will lead to double counting.

3.1.1 Methodology Development

In order to optimize the system for increased performance for curation a number of experiments were run in an attempt to minimize runtime; collision detection was by far the most computationally expensive area within the data curation process.

While the Levenshtein distance algorithm and longest common substring methods are in themselves able to identify most solutions in each instance the *Exact Duplicate* and *Begins With/End With* operators are implemented beforehand. This is due to the fact that these two operators are extremely cheap in regards to runtime and any collisions detected will save the system passing these to the more powerful yet slower detection methods, the correct ordering of comparator operators has a significant impact on runtime. Initial tests were run on a sample corpus of 1,326 new items; the success of each method in detecting collision is presented in table 3.1.

Table 3.1: Collision Detection Success Rate.

#	Detection Method	Collisions Found	Percentage Detected
1	Exact Duplicate	91	59.5%
2	Begins With/Ends With	98	64.1%
3	Longest Common Substring	145	94.8%
4	Levenshtein Distance Algorithm	152	99.3%
Total	153		

When examining the number of duplicates within the sample corpus 153 are detected when implementing each of the four methods in sequence, however as can be seen in table 3.1 all but 1 is detectable through usage of the Levenshtein Distance Algorithm alone it appears that one collision was detected through the longest common substring approach which the distance algorithm did not, while the longest common substring method identifies 94.8%. These results could be improved however by lowering the criteria by which a duplicate is identified, however as noted it is preferable that the system is rather conservative, favouring generating a greater number of false negatives than false positives. This extremely short and cheap comparison saves the need for passing such collisions to a more complex and expensive detection method, similarly a detection method based on the insertion of text at the beginning or end of a document detects a further 7 collisions.

A comparison of runtime for the different detection methods is presented in figure 3.1, this shows the run time in milliseconds when the four methods were run independently and then in sequence. The dataset used for this experiment contained 1,326 individual news items of which 153 were found to be duplicates and was used for each configuration to gauge the performance.

Examining the runtime of each detection method presented in figure 3.1 as would be

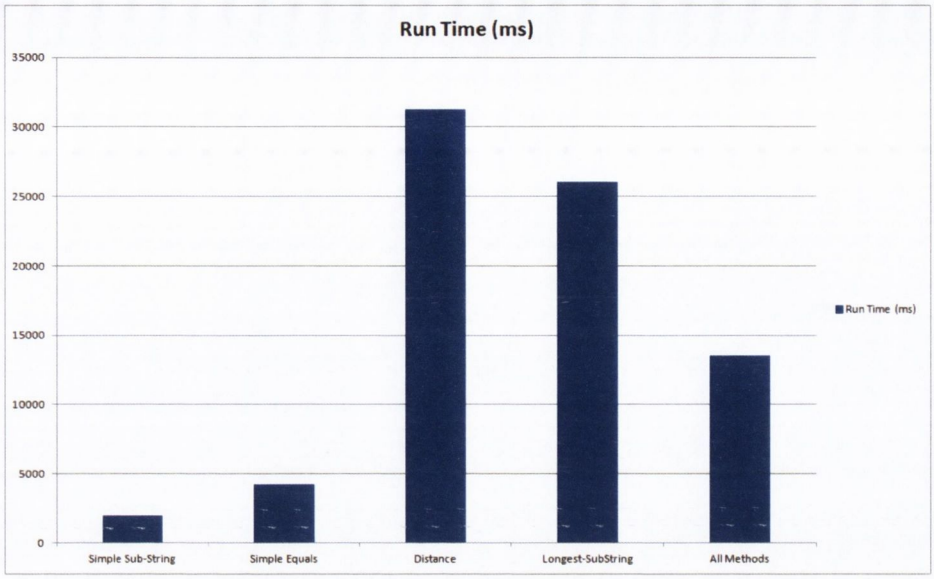


Figure 3.1: Runtime of Varying Collisions Detection Methods.
This shows the runtime for the collision detection methods. Values are presented in milliseconds, this is based on an evaluation of 1,326 news items, of which 153 were identified as duplicates.

expected the first two methods (Exact Equals, and Begins With Ends With) are extremely fast while still reducing the number of comparisons by a significant amount. Indeed when implementing all detection methods in a sequential order as presented in figure 3.2 there is a noticeable decrease in runtime when compared to running the distance or longest common substring algorithm independently. My strategy was to use all the four methods but in a sequence. Method 1 (exact) will identify 60%, method 2 (Begins/Ends with) picks up the next 5% of duplicates not identified by method 1. Method 3 (Longest common substring) catches 30% duplicates and, method 4 (Levenshtein Distance) around 5% more. My strategy leads to a reduction in processing time by using a mixture of expert detection methods. By applying a exact duplicate and begins with ends with before the Longest common substring and Levenshtein Distance I reduce the number of comparisons these two methods will carry out. By ordering them in this manner there is a 43% drop in runtime, when compared to the Levenshtein Distance approach (runtime was reduced from 31.3 seconds to 13.5).

3.2 Textual Analysis

Following is a discussion regarding the methodology of the extraction of textual information and the generation of a sentiment based time-series. The desire is that having developed a means for the collection, sanitisation, curation and if desired refinement of a corpus, to develop a means for a process by which a time series extracted from the corpus according to a given lexicon may be constructed. This affect time series in constructed by counting the frequency of

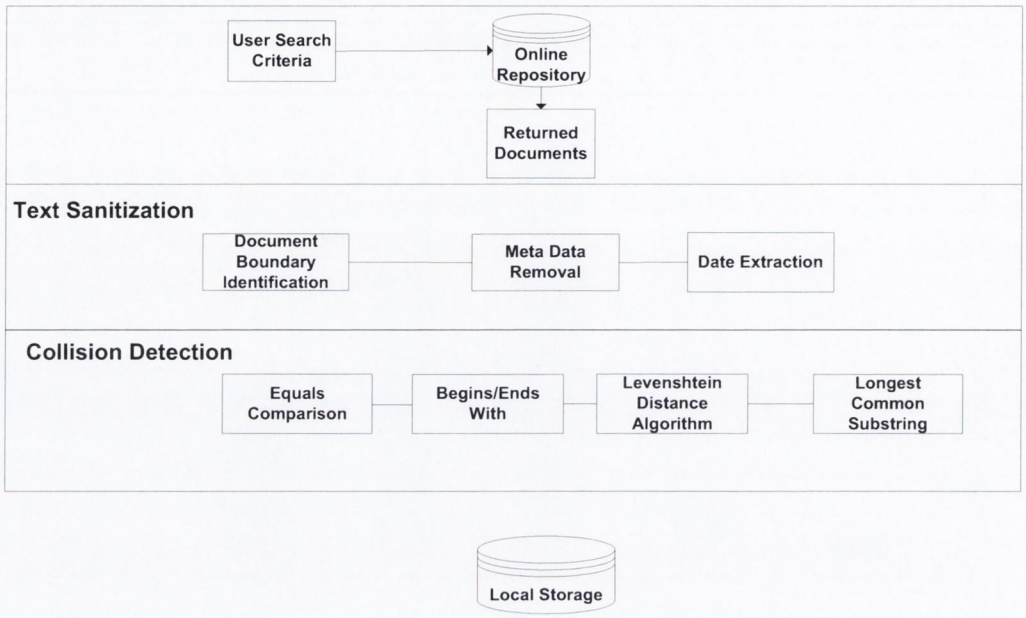


Figure 3.2: Data Curation Flowchart.

certain affect term which occur in news files on a given day.

3.2.1 Term Vector Construction

The method for the generation of affect time series implements the commonly used Bag-of-Words technique when combined with a reference legacy lexicon. Terms are evaluated individually from their surrounding context, and while this might prove simplistic it remains a commonly applied approach across many areas of textual analysis and has been demonstrated to perform well across a number of problem domains (section 2.3.2).

Within each document in the corpus the process identifies the occurrence and frequency of terms given within the legacy lexicon as being of affect nature; as the interest is within the areas of sentiment analysis these categories are generally *Positive* or *Negative* in nature. Using this lexicon a term vector is developed for each document as outlined in figure 3.3 recording the frequency of each individual term and storing total article length for the purpose of term weighting when applied.

While all terms at initially treated as equal assigned a value based on the raw frequency of the terms, a storage of the identified terms is of importance as term weighing functions are dependent both on the terms frequency within the document and its frequency across the corpus as a whole (section 2.3.2).

Considering two words which are found in general language *competition* and *foreign* according to the GI classification competition (as a noun) and foreign (as an adjective) both are given a negative evaluation, this assertion is probably true in that foreign beings being aliens,

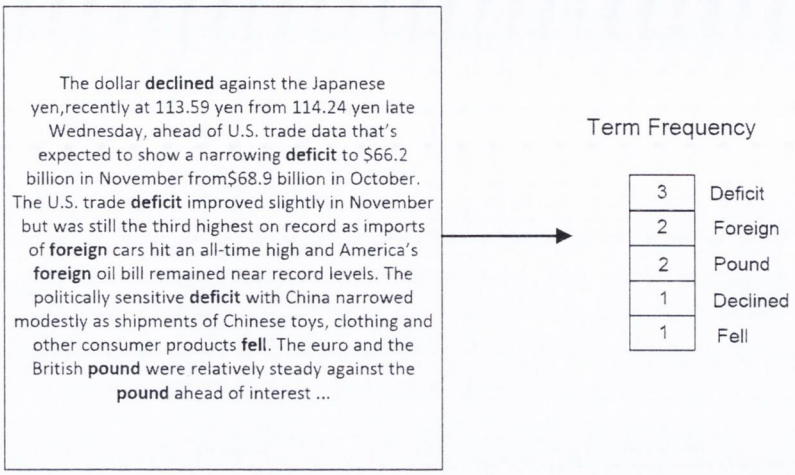


Figure 3.3: Term Frequency Vector Generation From General Inquirer.
The above figure outlines the construction of a document term vector of negative sentiment according to the General Inquirer.

hence threatening and competition relates to the invariable success of the strong. However in the language of finance the word foreign in a neutral sense- in terms of foreign investment and foreign exchange, and competition according to current economic views is regarded as something positive.

3.2.2 Affect Time Series Generation

As is generally the case where multiple articles occur within a single day each are combined, and examined as a single long text article. Evaluating each data point as a single large article demonstrates the impact which collisions may have on the analysis, the inclusion of duplicate news items (exact or near) would lead to a high degree of double counting artificially inflating any produced affect score. Each news item in the corpus has a date (of publication), the system extracts the key terms presented to it and notes the frequency for each of the term for everyday which can be referred to as the *sentiment score*. As there is more than one item of news generally published on the same topic on the same day my system aggregates the sentiment score in each document published on the same day and generates the sentiment score for that particular day.

My system iterates through each news item, stored in a news database, and by combining each article occurring on the same day into a single large text file assign a *Sentiment Score* to each day in the time series. With such a process in place variations in sentiment over time extracted from a provided corpus and lexicon can be examined such as that presented in figure 3.4. This time series is developed from negative affect as defined by the General Inquirer extracted from a corpus collected from LexisNexis regarding the United States economy;

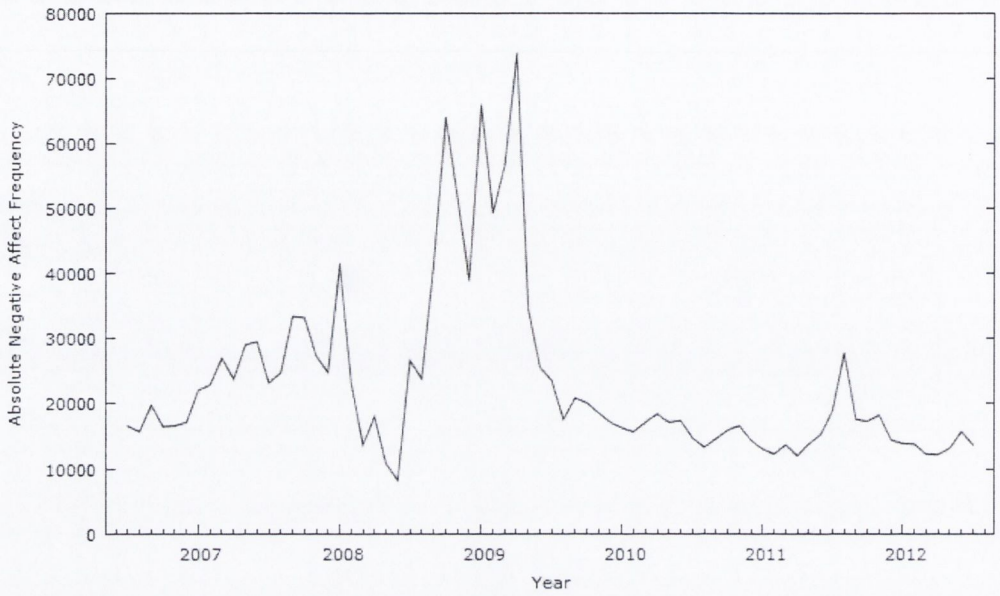


Figure 3.4: Negative Affect Time Series Compressed Monthly.

through this we are able to quantify such qualitative data towards the production of a quantitative time series. If our interest is within an examination of the impact of sentiment on finance markets such an affect time series allows for a comparison between these two time series, with one being based on technical financial changes over time, and the other based on extracted qualitative information. An examination of these two time series and any potential relationship forms a pivotal role in this research, whereby with this quantification and development of a numerical time series statistical analysis and comparison is possible over time. From which an affect based *Time Series* can be constructed.

3.2.3 Remarks

At this point the development of a methodological approach which allows for the ability to retrieve and curate large amounts of textual data has been presented; and from which extract qualitative data from a provided lexicon allowing for the development of a sentiment time series with minimal human intervention. The ability to examine the impact of collisions is provided through the system's ability to identify duplicates in an automated and consistent manner, however as the role played by duplicates is unknown it may be desirable to allow for the inclusion of duplicates. While previous works have aimed towards the development of systems for the identification of duplicate text items, to the best of my knowledge there has been no comprehensive examination on the frequency of such duplicates or the impact, but rather aimed towards their removal from datasets.

The goal is to expand this approach so as to provide it with a legacy lexicon and through

machine learning techniques refine such a lexicon to produce one which better captures sentiment within a given domain, in this instance the impact of sentiment within financial news. This refinement is achieved through the usage of two biologically inspired algorithms, the training and implementation of which will be covered following.

3.3 An Algorithmic Base For Dictionary Construction

Following is a description of the development of the methodology whereby a refinement of legacy lexica is produced through the usage of genetic algorithms and particle swarm optimization; whereby terms which within the given domain are misclassified are removed. As long as there is some method to measure the performance for textual analysis task domain specific lexica can be produced in an automated manner.

3.3.1 Evolutionary Algorithms

Following is an in depth analysis of the key issues in regards to the implementation of genetic algorithms and particle swarm optimisation, specifically within the area of feature selection.

Encoding

While a number of coding variations are available within GAs, the simplest allows for the binary encoding of a string of size *L*. Each bit is assigned a True or False value. Dependent on the problem domain these values may indicate the presence or absence of a feature or input. This is illustrated in figure 3.5 with a solution of length 10, where a 1 indicates true and a 0 indicates false.

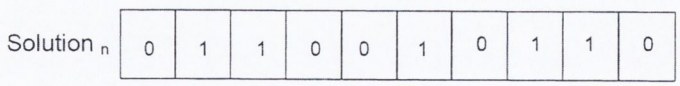


Figure 3.5: Example of GA Solution Encoding

When applying this method to lexicon refinement each feature in the solution would represent an individual term from the given lexicon. Terms which are 1 in a solution will be counted and used to calculate an affect score for that particular document. Features which are 0 represent a term which will not be counted and will play no role in the affect score calculation. The aim of this is that the system will produce a solution where terms which have no affect bearing in the domain are set to false and excluded when calculating an affect score. This issue is covered in greater detail in section 3.4.1

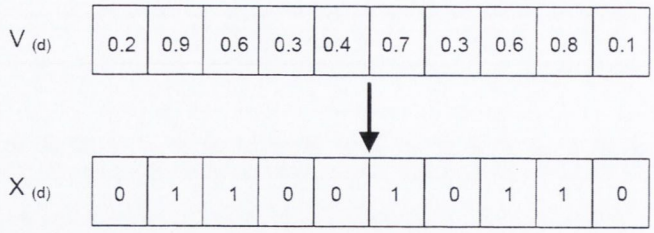


Figure 3.6: An example of Particle Swarm Encoding for Feature Selection

While genetic algorithms encode a candidate solution as a binary string, in particle swarms each candidate is described as a vector of real numbers within a bounded range representing the particle's location within the problem space. For numerical optimization problems these values are then used in an effort to solve the given problem.

However as the interest here is in feature selection it is required that any real numbers is mapped onto a binary domain to indicate the inclusion or exclusion of a given term. While the initial implementation of particle swarms for optimization was developed with such real numbers later work expanded to examine binary values [47]. Within this work a particle is restricted within a space between zero and 1, the value given represents the probability of the bit being 1; such that if a particle has a locational value 0.2 it has a 20% chance of being 1, or within the field of feature selection being set to be true. This is demonstrated in figure 3.6, where the initial vector of the particle V_i contains values bounded between 0 and 1, these values are evaluated and transformed to binary True/False values.

Fitness Function/Space

A limiting factor of genetic algorithms and particle swarm optimization is the requirement for the definition of an appropriate fitness function with which to evaluate individual solutions. Any such function must provide an adequate representation of the problem area yet commonly issues arise in regards to the complexity of fitness function. As this function will be evaluated a large number of times it will be computed for each member of the population on for each iteration. As such if the proposed fitness function is computationally expensive the runtime may become impractical. When examining complex engineering problems approximations may be implemented so as to drastically reduce run time yet such approximations must not be overly simplified so as to provide a solution which may be incompatible with real world situations. The fitness function for particle swarm optimization is calculated similarly.[1]

[1] Similarly to genetic algorithms upon each iteration each particle is evaluated and assigned a fitness value based on its current location within the problem space. This fitness value is used to guide each particle ideally towards the discovery of an optimal solution. While fitness within genetic algorithms influences the probability of a candidate's selection for future generations and breeding within particle swarms this is not that case, rather this fitness value guides the particle whereby each remembers its personal best location and that of the population as a

Propagation

The role of propagation within genetic algorithms should be such that the system encourages the exploration of a wide range of solutions across the problems feature space, such solutions should be spread across the scope of available solutions evenly and allowed to *mature* before selection intensity increases so as to encourage convergence. However beginning with a system in which selection pressure is set too low the system will be little more than a semi-guided random search method.

Particle swarms by contrast maintain a balance between the individuals own cognitive learning, and greater social interaction by learning as part of a group through the global exchange of information.

Selection

Within the field of genetic algorithms *Selection* refers to the determining of which members of the population will continue for future reproduction and evaluation of the problem space.

While a wide number of selection methods are available all follow a similar initial precepts; once all solutions have been evaluated a set number will be selected for future propagation the probability of which is dependent solely on relative performance of the solution against the proposed problem when compared to the population as a whole. When examining selection within GAs two key concepts to consider are *Loss of Diversity* p_d and *Selection Intensity*, where p_d may be considered the proportion of individuals not selected for continuation into the followed breeding stage.

By their very nature genetic algorithms aim to mimic evolutionary behaviour whereby candidates with poor performance fail to pass on their *"genetic material"* while those which perform well generate a number of offspring which due to their genetic material have a high probability of survival. However a rapid loss of diversity may have negative effects such that within a short number of generations the population will be composed of candidates with near identical genetic structure, limiting the probability of new more successful individuals emerging. Ideally the loss of diversity should be as low as possible due to the fact that a high level of p_d increases the risk of early convergence [8].

A roulette-wheel or proportional selection based approach is a common approach and the original selection method proposed by Holland. This method may be considered analogous to a simple roulette-wheel where all candidates are assigned a range on a roulette wheel proportional to their relative fitness; those which are selected will propagate and contribute to future generations passing on their "genetic material". A random number is selected with the range of 0 and 1; the solution whose range includes this value will then be selected. Solutions

whole.

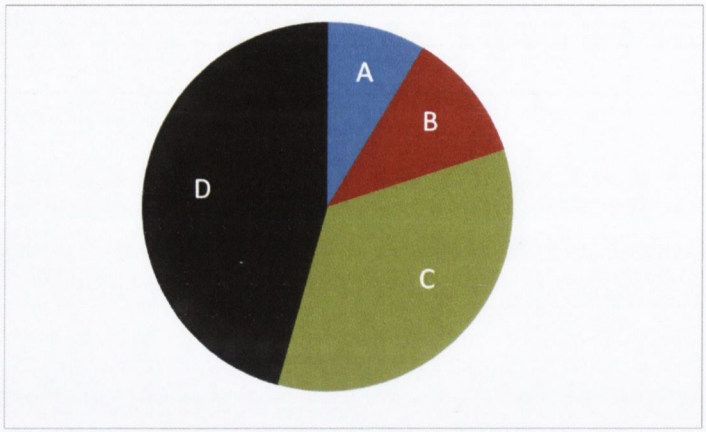


Figure 3.7: Roulette wheel selection method showing the relative performance of four solutions.

may be selected any number of times, however for the purpose of diversity it may be desired to exclude the breeding of one solution to another with an identical solution sequence as such pairing would simply result in the *cloning* of a solution [25].

An example of a Roulette Wheel based selection method with few candidate solutions is presented in figure 3.7. The size of the candidates range on the wheel is proportional to its relative performance. Solutions D and C can be seen to have performed well and as such have a high probability in regards to their continuation into future generations while solution A has performed poorly and as such has a low probability of selection for future propagation.

The simplicity of a roulette wheel based selection method allows for the simple calculation of a candidate's probability of selection for breeding and propagation. The probability of solution j being selected (p_j) is calculated by dividing the performance of the solution (f_j) by the sum by the sum of the fitness of all n solutions ($\sum_{i=0}^{n} f_i$).

$$p_j = \frac{f_j}{\sum_{i=0}^{n} f_i} \tag{3.1}$$

However in early generation so as to prevent early loss of diversity it may be desirable to implement methods which would encourage the survival of solutions even if their relative performance is rather poor. This is generally achieved through the usage of a scaling function whereby the end performance of a candidate will be artificially increased. Such scaling increases may be of a dynamic nature such that as more generations pass the selection intensity increased with the aim that given enough time to explore a greater area of the problem space the system then begins to encourage the population to converge, concentrating exploration on a smaller area where fitness peaks have been discovered. Criticisms aimed at a proportional selection method question the performance in terms of encouraging an appropriate level of

selection intensity and exploration. Arguments stem from the effect of common fitness scaling techniques which lead to little or no difference in selection probability between best and worst candidate solutions.

A common alternative selection method is a tournament based system. Such methods are simple to codify, a total of τ referred to as the tournament size)candidates are selected from the population at random. The candidates are selected at random from the population, their performance value has no impact on their probability of selection. The candidate with the highest performance within the selected pool to continues for future propagation. The tournament size τ can be modified over time, if τ is increased a greater number of candidates are selected for the tournament, this increases the selection intensity of the population. When examining the probability of an individual candidate being selected for continuation the analysis is more complex than that of proportional selection. Beginning with the assumption that all individual candidates are equally probable at being selected for tournament selection the probability of candidate j being selected for the tournament P_j may be calculated as

$$P_j = \frac{1}{N}\tau \tag{3.2}$$

Where N is total population size and τ is the tournament size, it can be seen that the probability of an individual candidate's selection is determined by the tournament size, as such modifications towards the size will have an impact towards loss of diversity. As noted the selection for a candidate to take part in a tournament is independent of the candidate's fitness value. As generally only a single candidate will emerge from each tournament the probability of survival is based on the proportion of candidates with a lower fitness value. Let $S(f_j)$ refer to the number of candidates within the population with a fitness value lower than or equal to f_j. Therefore the probability of candidate j being selected for the following generation may be calculated according to equation 3.2, from which the probability of a population whose performance is lower than that of candidate j may be calculated as

$$S^*(f_j) = N\left(\frac{S(f_j)}{N}\right)^{\tau} \tag{3.3}$$

Equation 3.3 shows the strong influence of the tournament size (τ) of a candidates probability of selection. A relatively small increase in the tournament size can result in a significant increase in selection intensity, if set too high in early generations the population may risk converging at a rapid rate. Combining equations 3.2 and 3.3 the probability of candidate j being selected and surviving (P_j^*) gives equation 3.4.

$$P_j^* = N\left(\frac{S(f_i)}{N}\right)^{\tau} * \left(\frac{1}{N}\right)\tau. \tag{3.4}$$

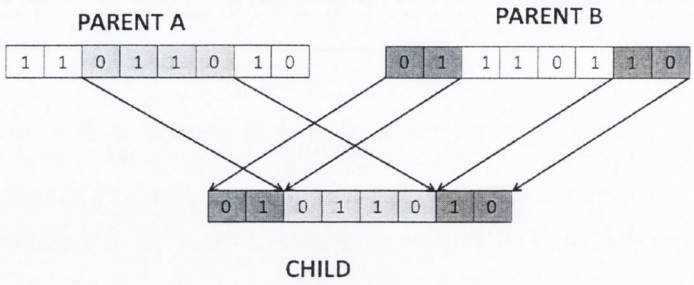


Figure 3.8: A Two-Point one to one Crossover.

Breeding

Once a suitable number of candidates have been selected for propagation the new generation of solutions may be generated. An appropriate breeding methodology is found to be a key issue within GAs, variations of the method for the generation of future candidates must encourage a wide exploration of the problem space while preventing the early destruction of high performing candidates through a high level of genetic interaction between individuals selected for breeding.

Various methods regarding the exact implementation of breeding have been proposed, following from the biologically inspired nature of GAs two parents are selected and some portion of each contributes towards the generation of a future offspring(s). The simplest method of breeding and that initially proposed by Holland concentrated on a 1-point crossover. A random point within the string α is selected, all bits ranging from $1 - \alpha$ from parent A are passed on to the following solution and bits $\alpha - L$ from parent B. This approach was later expanded to examine the impact of a two-point crossover implementation illustrated in figure 3.8. From this researchers have expanded to a n point crossover whereby a random number n cut points are selected. Uniform crossover is a slight variation of an n-point crossover where each bit is randomly selected from a parent according to a probability $P_s = 0.5$, this was later expanded upon [98] so as to use selection bias where P_s may be a value other than 0.5. A crossover method where the parameter P_s may be any value gives the advantage of allowing for a variation of the level of interaction between candidates over time. Varying the P_s allows for a greater level of genetic material from one individual to contribute towards offspring, thus allowing for a greater level of preservation from well performing candidates structure yet allowing for the introduction of variations for the purpose of problem space exploration.

Mutation

In its simplest form mutation within GAs results in a small number (general less than 1%) of bits within the population being flipped, the inspiration of this is so as to maintain a level of diversity within the population. If selection intensity was set at too high a rate it is possible

that from an early stage a single feature of the problem space may be left unexplored. Consider the population :

$$A_1 = 011110$$
$$A_2 = 110010$$

Through crossover alone the final bit may never be set to true since the i_{th} bit of a child solution must be the same as the i_{th} bit of one of the parent solutions, as a result this feature will be excluded from all future analysis. Mutation aims to introduce an aspect of diversity within the system to help prevent premature convergence resulting in a sub-optimal local maxima being returned.

A long running debate within GAs is the relative importance and impact of mutation and crossover. Researchers have developed the crossover mechanism employing adaptive crossover rater where one parent will play a greater role than the other in the crossover mechanism. By varying this crossover rate over based on the number of generations periods of exploration are encourage early on while towards the end the population converges. Using this adaptive configuration has been shown to improve results in complex tasks [7, 91].

While crossover may result is high levels of information exchange between two parents this may also result in a high level of information destruction whereby *"the behaviour of each offspring becomes only minimally related to the behaviour of the parents"* [28]. Mutation by contrast will guarantee is low level of variation of candidates from one generation to the next by definition, and while such implementation may be viewed as a purely random search of the feature space, selection will ensure that the entire process will not be based on a purely random search and that through the competitive nature of selection convergence is guaranteed. However later studies [92, 97] have questioned this view, arguing that crossover plays a far greater role in exploration within genetic algorithms, and that that the exchange of information between two or more members of the population is a key component. Indeed, by varying the level of interaction between parents it is possible to ensure that the level of information exchange and subsequent destruction is minimized. These works propose that the main and perhaps sole purpose of the mutation operator is to allow for a means of discouraging early convergence through the inclusion of a certain level of randomness to the system. While a number of works have attempted to provide guidelines regarding crossover and mutation rates [35, 36] with varying rates of success further works have proposed that these configurations are not guaranteed to perform well across all problems and that system configuration may be specific to the particular problem. As the configuration of the system can have such a large impact on the performance researchers have proposed that system configuration is an optimization problem in its own right.

Diversity and Convergence

As noted a key area of research towards propagation within GAs is towards maintaining an acceptable level of diversity within the population.

Initial candidates should be evenly distributed across the problem feature space thus allowing for adequate exploration of all regions, as time passes candidates will begin to migrate towards various region and groups will form. Within later generations the goal would be such that as the population matures candidates converge towards what should be a global maxima in the problem space. However if system convergence occurs at too early a rate, candidates may group towards *Local Maxima* failing to explore further regions and resulting in a sub-optimal solution.

Configuration of selection and recombination play key roles in the rate of system convergence and their varying effect has been a key area of interest towards research in GAs. Much research has shown that ideally the selection intensity of the population should vary over time thus allowing for the prevention of early convergence.

A number of means for the measurement of diversity regarding a population have been proposed; within a simple binary encoding representation the Hamming Distance which measures the differences between two binary strings can be used. Alternative measures examine the percentage of solutions with no exact duplicates [52], however as noted if the encoding length L is sufficiently large even a very low mutation rate will cause a high number of flips within the population and thus would suggest a low level of exact duplicates, and do not consider how similar or different solutions are which would provide a more accurate representation regarding the diversity of the population. However given its simplicity to calculate it does provide some measure of the diversity of the population and if this diversity measurement is low then *"any other measure of diversity will also have a low value"* [24].

Alternative methods have been developed in which candidates within close proximity to one another *share* their fitness value, while allowing the exploration of a particular area of the search space discourage concentration and grouping across a small area [41]. Clearly any methods developed towards the maintaining of diversity would towards later generations lessen the exploration encouragement as without this the system will not be allowed to converge towards a global maxima, becoming analogous to a guided random search method.

3.4 Lexica Refinement

The goal of this work is to examine lexicon refinement through genetic algorithms and particle swarm optimization as applied in a feature selection manner; each term within a given lexicon is a feature within the problem space the impact of the inclusion of which is evaluated in each

iteration. The resultant solutions in each instance will produce a Boolean vector indicating the inclusion and exclusion of a term in regards to the aggregation of an affect time series. While this method should be applicable across a range of text classification domains this work aims to evaluate these methods within the field of sentiment analysis of financial news.

The two methods for feature selection, genetic algorithms and a binary particle swarm are applied towards the given corpus against a market index for use as a training platform. While the initial examination regarding an improvement (*if any*) is of interest as the goal is to improve sentiment analysis within financial news a secondary output for examination is the refined reference lexica. Through an examination of terms identified for exclusion and inclusion by the system this provides the potential for gaining a greater insight into the language found within financial news.

In order to train the system to produce a refined lexicon for then index the system must be given a training dataset. Six months of news items and market returns were used, the system used regression analysis of the generated affect time series, with the R^2 value used as the performance measurement which was obtained from the regression model discussed in section 3.4.2. The R^2 value was used as it gives a measurement of the explanatory ability of the regression models for market returns. Once the training process was completed a refined lexicon was produced, the affect time series which was generated from this subset of the original lexicon best described movements in the market returns timeseries. As this time period is used for training of the system only, this portion of the time period will not be included in later evaluation of results. This is to account for the fact that results risk being the result of data mining and overfitting rather than showing a true relationship between a refined lexicon and market behaviour.

3.4.1 Feature Space Definition

If a system is provided with an input of a legacy lexicon a problem space may be defined towards feature selection methods. Candidate solutions are created the size of which is the number of terms in the lexicon, in both genetic algorithms and particle swarm optimization all initial candidates are randomly generated. Each of these candidate solutions represents the terms to include and exclude when calculating a sentiment score for the document being evaluated. Initially these solutions will be random when it comes to which terms to include and exclude when calculating an affect score. This is illustrated in figure 3.9, where a document has been collapsed into a term frequency vector D_j with the frequency of terms recorded $W_{1,i}$. This is processed against the candidate solution n; within the solution a 0 indicates the term is to be excluded, while a 1 indicated that the term may be of relevance and its frequency within the document will be included in any affect score calculated. In this instance the frequency of

terms 1, 4, 5, 7 and n will be ignored and proposed as being of no interest in the given domain.

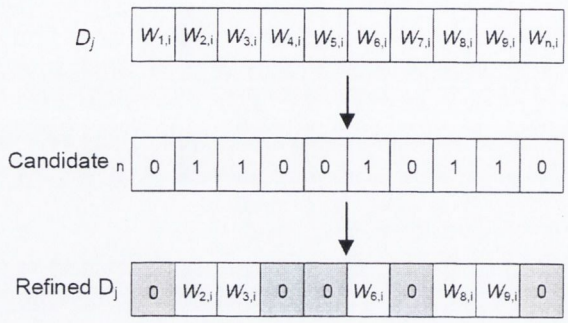


Figure 3.9: Lexicon Refinement

3.4.2 Evaluation Benchmark

Within genetic algorithms and particle swarm optimization some form of evaluating the performance of the system in an automated manner is needed so as to guide and gauge improvements in the system. In the field of textual analysis there remain a number of appropriate evaluation functions, such as the accuracy of classification of pre-annotated documents. My interest here is the improvement of textual analysis for financial news and as has previously been attempted use the markets themselves as a training benchmark [51]. If there is indeed a relationship between sentiment in news and market movements an examination of the resulting affect time series against market behaviour should provide a proxy in regards to the accuracy of our refined lexicon. In sentiment analysis this may evaluate a company's market movements against sentiment in firm specific news, or examining market index movements against general financial news [104]. When examining financial indices rather than examining the closing price or change of an index it is far more common to examine the returns of prices. Returns are used rather than prices, prices between consecutive days are highly correlated while consecutive price changes are not [101]. Returns are also used over a raw change in price is based on the fact log normality where price changes are normally distributed (this may or may not be true) and is better suited for a number of financial models. Returns of market prices are also concentrated on as these are used in previous studies of sentiment analysis of financial news, many of these having input from economists with expertise in the area of finance. The return of a financial instrument R_t is the natural logarithm of its price P_t on two consecutive intervals when the trading was carried out; for stock prices one usually deals with closing prices on two consecutive days as calculated in equation 3.5.

$$R_t = ln\left(\frac{P_t}{P_{t-1}}\right) \tag{3.5}$$

An analysis of returns provides a greater insight into market behaviour that an examination on prices or changes [101], and unlike closing prices and changes returns do not exhibit autocorrelation to the same extent. If closing price of a index are examined there will be a high level of autocorrelation as price changes on a day to day basis are small, is problematic when evaluating the results of a regression model. An in-depth analysis and development of statistical modeling of financial returns is outside the domain of this work, however previous researchers have proposed a number of means for the examination of sentiment against market returns. Here an approach similar to that implemented by Tetlock is examined; an affect time series is regressed against returns of closing market prices according to equation 3.6. The affect time series records the sentiment score for each day by evaluating all documents in a corpus and counting the frequency of a given list of terms, these values are orders in a chronological manner and a time series can be constructed.

$$Returns_t = \alpha 1 + \beta 1.L5(Returns_t) + \gamma 1.L5(Sent_t) \tag{3.6}$$

Where *Sent* is the calculated affect score for a given day and *Returns* is the market returns at time t, $L5$ is a lagged operator which includes a vector of five previous values of returns and sentiment score in that time series, such that $L5(x_t) = [x_{t-1}\ x_{t-2}\ x_{t-3}\ x_{t-4}\ x_{t-5}]$, lagged values are included so as to examine the potential delay in the impact of sentiment. Where α is a constant and the β value shows the impact of previous market returns. The values for γ will give a measurement for the impact of a change of one standard deviation of the sentiment measurement on market returns, describing the dependency of market movements on the sentiment measurement. In order to prevent any future knowledge sentiment from the previous day only is examined against the market returns as the time at which ca given news item was released can not be guaranteed. This model provides a valuable resource as it was developed specifically with an examination of sentiment on financial returns by individuals with expert knowledge within the field of financial analysis. This is however a highly simplified version of the model put forward by Tetlock, as he incorporates a number of control variables based on an in-depth understanding of the field.

From this regression model the resulting coefficient of determination R^2 may be examined, this value provides some level to gauge the predictive value of the produced model. R^2 indicates how well data point fit the model, giving the amount of movements that can be explained, taking into account the errors in prediction and the variability of the dataset. If y_i is the observed value in dataset y at point i, and $\bar{y}$ is the mean of the dataset calculated using equation 3.7.

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i \tag{3.7}$$

From this the total sum of squares can be calculated (SS_{tot}) to explain the variability in the dataset but summing the difference between an observation and the mean by using equation 3.8.

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.8}$$

So as to examine the predictive ability of the regression model the sum of squared residuals (SS_{res}) can be calculated. If f_i is the predicted value of y at point i this can be compared to the observed value y_i and the sum of the errors can be calculated by equation 3.9. This provides a measurement of how accurate the model predicts values in the dataset and the level to which it explains movements.

$$SS_{res} = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{3.9}$$

The coefficient of determination (R^2) can then be calculated using formula 3.10. This measurement takes into account the level that values move around the mean and the sum of the errors in the prediction of values in the dataset and can be used as a measurement of how well the regression model predicts or explains the dataset y.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.10}$$

This measurement is also chosen since it was also a key result given in previous work on sentiment analysis of financial texts [64]. Through the usage of R^2 can be assigned as a fitness value to the candidate solutions so as to guide and improve over time allowing for an appropriate lexicon development. Using R^2 as the fitness value will favour solutions which best describe the returns time series, this should guide the system to produce a lexicon containing the terms which best describe negative and positive affect within the domain for financial texts.

3.4.3 Methodological Implementation

Initially all news items are collapsed into a term frequency vector as shown in figure 3.3 using a legacy reference lexicon such as the General Inquirer negative affect category. Rather than combine affect terms to produce a single affect score for a document all terms are stored individually; firstly for the purpose of term weighting calculation and secondly to exclude terms as the system progresses.

A time series is constructed composed of data point on days where there is both market data and some news items have been returned. As is the case with both genetic algorithms and particle swarm optimization solutions are randomly created across the problem space, each candidate solution is a vector of length L where L is the size of the given reference lexicon. Once generated a time series is developed from each of the proposed solutions, if a term is

encountered which is set to 1 within the given solution its frequency is counted and summed with the frequency of all other affect terms occurring on the given date. The system is given six months' worth of news and market data with which the system regresses the affect time series against a given market index return time series using formula 3.6. The resulting R^2 value is set as the solution's fitness value; the system process is repeated until a set criterion is met regarding the number of iterations. Once an optimal solution has been generated this new refined lexicon is used to produce an affect time series against the remaining textual data, so as to allow for comparison against the *Raw* lexicon in terms of system performance this time series obtained through this lexicon is also output.

Feature Selection:Genetic Algorithms

As discussed the binary/Boolean based nature of a candidate solution are well suited for the purpose of feature selection. Given that lexica tend to include a large number of terms which result in a large feature space care had to be taken so as to develop a system capable of exploring the problem space sufficiently. Many of the system configurations for genetic algorithms provided by Moser & Narasimha [71] were followed, rather than relying on fixed values regarding selection intensity, crossover and mutation rates all were dynamically altered based on the current generation of the system. However unlike Moser & Narasimha rather than implementing a roulette wheel based selection method, a tournament selection method was implemented, this was implemented as experiments and past studies have demonstrated that tournament selection outperforms roulette wheel in regards to exploration of the problem space [8].

The configuration of the key variables for the genetic algorithm method are presented in table 3.2. Through the variation in crossover rate and bias the intensity of the system towards convergence is controlled. Crossover rate refers to the percentage of solutions to which, once selected, a crossover function will be applied. In effect it allows of a percentage of solutions to proceed unaltered. A high crossover bias allows for a greater level of one solution to be passed on when compared to the other, after both have been selected for breeding. This allows for periods of exploration in generations followed by periods where selection intensity becomes greater leading to convergence of the population.

The average fitness of the population as a whole can be examined in figure 3.10, this imagine illustrates the average fitness of the population of a typical evaluation over time. During the initial period of *exploration* (generations 1-50) there is a modest improvement in average fitness over time, with a low level of selection intensity exploration of the problem space is encouraged. In the second stage *growth* (generations 51-100) selection intensity increases where solutions begin to concentrate towards feature spaces with a high fitness value, this leads to an increase in the rate of average fitness improvement. In the third stage *maturation* (generations 101-150)

Algorithm 1: Genetic Algorithm Lexicon Refinement.

```
Input  : Reference Lexicon of size L
Output : Refined Contemporary Lexicon
Data: Time Series of Affect Term Vectors, Time Series of Market Index
Result: Lexicon Refinement Through Genetic Algorithm
Initialize a population of size N solutions;
while Current Generation is Less than 100 do
    i=0;
    j=0;
    forall the Solutions do
        forall the Documents do
            Evaluate Document[j] Term Vector Against Solution[i];
            if Term=True in Solution then
                Retrieve Term Frequency in Document;
                Increment Affect Time Series Datapoint ;
            else
                Ignore Term;
            j=j+1;
        Calculate R^2 Value of Affect Time Series against Returns Time Series;
        Assign Fitness Value (R^2) to Solution[i];
        i=i+1;
    Implement Selection Method;
    Implement Crossover;
    Current Generation=Current Generation+1;
Generate Time Series on Remaining Corpus Data Using Refined Lexicon;
```

selection intensity increases further, as convergence occurs the rate of improvement of average performance decreases as solutions have gathered around a smaller area of the feature space. By the time the system has entered the final stage *convergence* (generations 151 onward) there is no improvement in the average performance as the population has completely converged, with little diversity within the population further exploration is unlikely, at this point the system will terminate a produce the optimal solution discovered. These changes in behaviour are caused by the dynamic configuration of the system as shown in table 3.2, the inflection points at the 50th, 100th and 150th generation indicate periods where the selection intensity increases and the level as does the level of crossover. If selection intensity is constant and remains low for all generations the increases in performance are more moderate, however the end solution may be be sub-optimal as the population does not converge to the same extent. If however the selection intensity is constant and set too high the increase will be rapid and the system will converge in a shorter period of time but again my produce a sub-optimal solution. This is in contrast to the

Table 3.2: Genetic Algorithm System Configuration.

Age	Parameter	Value	Effect
0-50	Tournament Size	2	Moderate selection pressure
Exploration	Mutation Rate	2-3%	Focus on Exploration
	Crossover Rate	75%	Moderate Chromosome Interaction
	Crossover Bias	90%	Low Gene Interaction
50-75	Tournament Size	3	
Growth	Mutation Rate	1%	Exploration and combination
	Crossover Rate	80%	
	Crossover Bias	75%	Moderate Gene Interaction
75-100	Tournament Size	4	
Maturation	Mutation Rate	0.50%	Focus on maturing
	Crossover Rate	90%	Increased Chromosome Interaction
	Crossover Bias	60%	Increased Gene Interaction
100+	Tournament Size	4	
Convergence	Mutation Rate	0%	Enforce Convergence
	Crossover Rate	100%	High Chromosome Interaction
	Crossover Bias	50%	High Gene Interaction

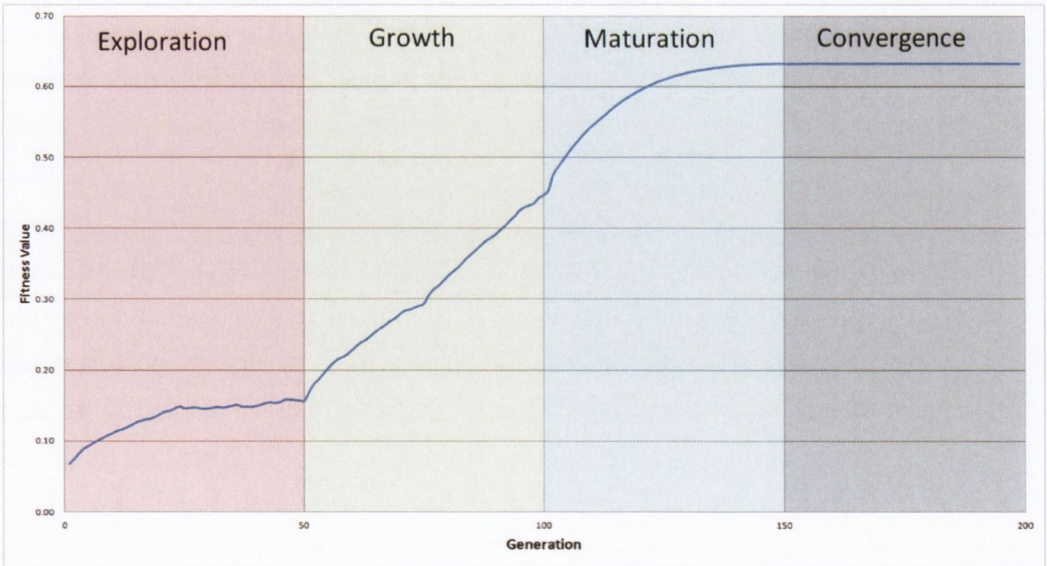


Figure 3.10: Average Performance over Time Through Genetic Algorithms.

increase in average fitness seen in figure 3.11, within particle swarm optimisation there are no variables which impact on the convergence of the system. The increase of fitness is gentler and since the configuration is constant there are no inflection point seen within this system. In the case of PSO the population seems to converge at a faster rate, after approximately 80 iterations compered to 150 in genetic algorithms.

Feature Selection:Particle Swarm Optimization

While in many ways rather different in nature the underlying features and implementation of particle swarm optimization for feature selection are rather similar. Upon initialization each particle is randomly spread throughout the problem space, their locational values are bounded between the values 0 and 1, as noted while originally intended for real value usage particle swarms may be applied to binary problems such as feature selection. V_{id} is a vector of size n where n is the number of features which represents the velocity of the particle. V_{id} is a continuous value bounded between [0,1]. Recall that $X_i d$ refers to the location of particle i within dimension d (see section 2.5.2). While a particle can be located anywhere within the problem space with a value between 0 and 1 when calculating position afterwards it is mapped to a binary value. By comparing the value of V_{id} against a random number between 0 and 1 the position X_{id} can be mapped to a binary value across each of the dimensions a seen in equation 3.11.

$$X_{id} = \begin{cases} 1, if & rand() \leq S(V_i d) \\ 0, if & rand() > S(V_i d) \end{cases} \tag{3.11}$$

Where $S(V_i d)$ is a sigmoid limiting transformation which maps $X_i d$ to the values 0 and 1, $rand()$ is a random number between 0 and 1 [47]. In this manner particle swarms can be used for lexicon term reduction similarly to genetic algorithms. Solutions are evaluated and an affect time series is constructed according to the given solution. This time series is regressed against the market returns and the R^2 value assigned as the fitness value. An advantage of particle swarm optimization is that within its simplest form little must be decided in terms of variable selection, as there is no crossover, selection or mutation operator. Both genetic algorithms and particle swarm optimization emphasis random selection. However the performance of a proposed solution measured by in terms of the R^2 value is in a sense used to guide the solution. When examining the average fitness at each iteration there is an upward trend where the average fitness and performance of the population increases over time (see Figure 3.11). As the particles are motivated by the location of both their own personal best and the global best location they are encouraged to explore these areas in greater detail. Unlike the average fitness seen in the GAs the improvement over time here is smoother, this is caused in part since the configuration of the system is static. There are no variables which are hard coded to change at

Algorithm 2: Particle Swarm Optimization Lexicon Refinement.

```
Input  : Reference Lexicon of size L
Output : Refined Contemporary Lexicon
Data: Time Series of Affect Term Vectors, Time Series of Market Index
Result: Lexicon Refinement Through Particle Swarm Optimization
Initialize a population of size N Particles;
Distribute Across Problem Space;
Personal Best=0;
Global Best=0;
while Current Iteration is Less than 100 do
    i=0;
    j=0;
    forall the Particles do
        Map Particle Location to Binary Values;
        Evaluate Document[j] Term Vector Against Particle[i];
        forall the Documents do
            if Term=True in Particle[i] then
                Retrieve Term Frequency in Document[j];
                Increment Affect Time Series Datapoint ;
            else
                Ignore Term;
            j=j+1;
        Calculate Velocity;
        Update Position of Particle[i];
        Map to Binary Space;
        Calculate R^2 Value of Affect Time Series against Returns Time Series;
        Performance=R^2;
        if Performance (R^2)≥ Personal Best then
            Personal Best=Performance
        if Performance (R^2)≥ Global Best then
            Global Best=Performance
        i=i+1;
    Current Iteration=Current Iteration+1;
Generate Time Series on Remaining Corpus Data Using Refined Lexicon;
```

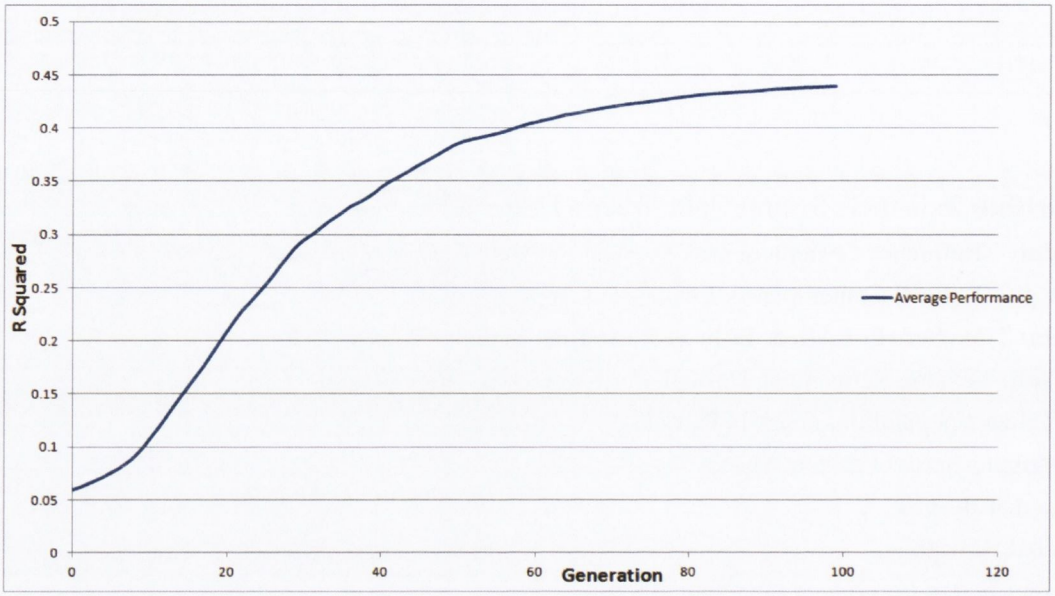


Figure 3.11: Average Performance over Time Through Particle Swarm Optimization.

set times which result in an increase in selection intensity. This continues until after a certain number of generations convergence occurs whereby most particles have centred themselves within a small region within the problem space.

3.5 Chapter Summary

This chapter has outlined the methodological development to test the ability to develop a refined contemporary lexicon from a given reference legacy lexicon in an automated manner.

Firstly the issue of data curation was discussed, the large scale collection of a corpus and examination of its contents. In the corpus a significant number of collisions were found, news articles were re-released at times with no modifications and others where a significant amount of content has been modified. The identification of such collisions and development of a system to identify such collisions in a reliable manner was in itself a significant task. From this collected, refined and archive texts and corpus generation the system was developed towards the generation of an affect based time series, which through the usage of a given lexicon provides the ability to track changes of sentiment within a given corpus as a time series. This allows for the examination of sentiment in a statistical and quantitative manner; examining how sentiment changes over time with peaks and dips. The development of a methodology was outlined in which a given reference lexicon could be refined, with terms removed which within a presented domain provide little or indeed detrimental information due to misclassification within the domain of interest.

Chapter 4

Case Studies & Evaluation

4.1 Introduction

The aims of this chapter are to examine the role sentiment may play in future movements within financial markets, the examination of a refined lexica and contemporary lexica in regards to capturing the *mood* of the market while potentially improving analytical results. The examination of a choice of words, or the solution in genetic algorithms and particle swarm optimization, which best indirectly captures the mood of the market. For us having a regression equation of prices, which comprises sentiment scores also which produce the best estimation of future prices as measured by R squared.

Three key areas of interest are:

1. The Role of Duplicates within News Analysis.

2. The impact of different legacy reference lexica in sentiment analysis.

3. An examination of the refined lexica produced by the proposed methodology.

Tetlock has claimed that: (a) he has constructed a "simple measure of media pessimism from the contents of the WSJ('Wall Street Journal') column", and (b) has estimated inter temporal links between "this measure of media pessimism and the stock market using vector autoregression" [104](page 1140). The media pessimism measurement was constructed by analysis a news-cum-opinion column in the WSJ, called *Abreast of the Market*(AOTM): 'AOTM is one of the most widely read market summary columns in the United States. It provides analysis of prior market activity, describes some notable company-specific events, and sometimes offers predictions for the future' [22]. Tetlock analysed the column and Dow Jones Industrial Average during a rather dormant period 1984-1999, except for the 1987 crash-a total of 3,709 articles, however a dummy variable is included in analysis so that results are

not driven by this single observation. More recent studies of the impact of sentiment on stock returns follow a similar line in selecting summaries of financial news from WJS(AOTM), to investigate the bias of journalists writing the column (from 1970 to 1984, a total of 9,552 articles [22]) for investigating the differential effects of negative and positive sentiment during economic expansions and recessions. Both Dougal et al. [22] and Garcia [30] , DJIA is used as a measure of stock returns: Tetlock used the General Inquirer and the other two use Loughran and McDonald's [64] list of *'financially'* relevant positive and negative words. Dougal et al. focus on the journalists who produce the column and as such do not use estimates of affect via word count, rather for them it is the journalist who is the proxy of sentiment they relate "average returns to the day a particular journalist writes".

It has been argued within the context of the US economy that AOTM has deliberately published misleading information about certain stocks in the 1920's [111]. Indeed, Dougal et al. show that journalists have a negative or positive bias that reflects in their columns, and further claim that this has an impact on the performance of DJIA. Perhaps news-cum-opinion articles are influential in determining part of the performance of stock returns.

The focus is on news articles released by the AP Financial Newswire that contain terms related to *economy* or *economic* using an annotated digital library of news articles available from the online repository *LexisNexis*. The annotation is carried out by LexisNexis semi-automatically, keywords are extracted from each news item, and statistically relevant keywords are then annotated digitally with the news item. Search algorithms then match the annotations to retrieve (statistically) 'relevant' news items. The system also has a classification of subjects; ranging from company activities, crime, economy and economic indicators to trade & development, and industry classification that included aerospace & defense to banking & finance, and from retail sales to travel and hospitality. News items can be grouped geographically and based on the usage of an individual company's name. LexisNexis can create a corpus of news items, each with individual date stamps, from the archive which comprises over 1,500 newspapers published across the US and world. The'United States' is used as the geographic region and two keywords 'economy' and 'economic' are used to retrieve news items. Except for the geographic reasons the category classification has not been used. There are two biases in this otherwise quasi-random data-set; first, the choice of categories in that news items that do not contain the words/phrases chosen by LexisNexis annotators but contain semantically related words/phrases will be excluded, second the choice of news publications, this is not compulsory, but have chosen AP Financial Newswire and that perhaps biases the choice of news source. The digital news archives do contain *duplicates*; news items that are either exactly the same word for word or contain variants of the same story. The duplicates could be corrections and updates; these variants of the same story are important as a news story evolves from unexpected happenings. Sometimes, though news could be repeated in a word for

word manner, by accident or design to reinforce a message. If it is by accident then a keyword or affect word can be over-counted and given one basis point change in the frequency of a negative affect word can cause more than 4 basis point drop in future stock returns [104], it is important to account for exact or near exact duplicates.

In Tetlock and others, the DJIA returns and detrended volume of stocks is used as a measurement of movement in NYSE: the volatility in this index is taken into account by doing a series of transformations on the residual of the DJIA; Dougal et al. and Garcia use heteroskedastic estimates of volatility by estimating GARCH models. Dougal et al. focus not on word counts but on the so-called journalist effect and have concluded that the analysis of S&P 500 returns has the same result. If the market moves unexpectedly then perhaps one might use indices like VIX - the square root of the risk neutral expectation of the S&P 500 variance over the next 30 calendar days. VIX has been popularly referred to as the *'Fear Index'*; however as it is an aggregate of a range of options on the S&P 500 perhaps VIX captures traders' sentiment as well as the sentiment extracted from news reports and opinions about the market. As the interest is in examining any potential link between sentiment extracted from financial news and future market movements experiments similar to those put forward by Tetlock and others who have provided extensive research within this field will be used. Similarly to previous works sentiment extracted from news which has be indicated as concerning the US economy will be examined against major financial indices within the United States. The three Indices for examination; Dow Jones Industrial Average, S&P 500 and VIX are selected as they represent key indicators of financial performance within the US and have also been previously studied in regards to the impact of sentiment on financial indices. In order to evaluate the performance of the system in regards to examining the impact of sentiment within financial news towards future market performance the impact of sentiment will be examine in a similar manner to Tetlock [104], namely a simplified version of the regression model put forward (section 3.4.2).[1] The results from the regression analysis aim to examine the previously identified observations within sentiment analysis and its impact on finance movements whereby a lag in regards to sentiment has been found to have a statistically significant impact on market returns, followed by a reversal, which has been argued to be a return to market fundamentals. Secondly the performance of refined legacy lexica once feature selection methods have been applied with the aim of the removal of terms which within the given context have been misclassified is examined. From this an examination of the terms selected for inclusion and exclusion within the refined lexica is also of interest as it allows us to examine firstly if terms which have been identified through manual examination are misclassified within this domain while also potentially providing insight regarding nuances

[1] Tetlock and others provide complex regression models including a number of control variables which are outside the scope of this research.

within the language.

4.2 DataSet

For an examination of sentiment within the domain of financial news there are two key aims in regards to data curation and corpus development; firstly an extensive corpus with a high number of articles occurring on a daily basis, while secondly as little human intervention in terms of the corpus selection and refinement is preferable. Articles are retrieved through the online repository LexisNexis, this repository provides the ability to search across a large number of sources with provided search terms and time frame refinement. A single source is selected for news collection, the Associated Press financial Newswire, this source provides an extensive number of articles on a daily basis and is frequently reprinted across a number of other news sources. The selection of a single source is to account for the event in which each source will be discussing a single event and as such should provide no new information however is a potential source of future research.

Lexica

In the GI dictionary the negative entries comprised nouns, adjectives, verbs, and, adverbs some of the entries just lemma while others have inflected or derivational forms as well, the total number of entries were 2,004. According to Loughran & McDonald [17] the coverage of the GI dictionary was not comprehensive on two counts; first not all the inflected or derivation terms of sentiment words were included and the authors show that upon including the different morphological form of entries the number of negative tokens in the GI dictionary can be increased to 4,187, this expanded word list (referred to here as Negative-Inf) was produced by Loughran & McDondald which inflected version of word "to forms that retain the original meaning of the root word" [64]. Second Loughran & McDonald also identified a number of words used in the financial language, which have a negative and positive connotation the negative terms amount to 2,337 and the positive terms to 353 (summarised in table 4.1). Loughran & McDonald claim that their contribution is not just that they have added more sentiment terms and variance but more importantly that 73% of GI negative entries are classified as negative but those words are used as terms in financial negative.

4.2.1 Corpus Collection

For the purpose of corpus development one consisting of a large number of articles with as few data points missing as possible is preferable. The criteria for article selection is intentionally left relaxed, articles must have been identified by LexisNexis as concertinaing the United States,

Table 4.1: Summary of Lexica.

Category	# Terms	Source	Domain
Negative	2,004	General Inquirer	General Language
Negative-Inf	4,187	General Inquirer	General Language
Positive	1,634	General Inquirer	General Language
Negative Financial	2,337	Loughran & McDonald	Financial

containing the terms *Economy* or *Economic* a minimum of three times, all articles containing less than a total of 15 tokens are excluded as such articles tend to provide simply tables or references to graphs which give no textual information. However as the aim is to limit the level of human interaction no further user based refinement is implemented.

The time-frame for selection begins in 2006 extending until August 2012, this time period provides a certain level of variation over time periods. During this time period there were major changes in the market with nearly all financial instruments having price inflation from 2006-2007 followed by losses in 2008, followed by modest recovery. Table 4.2 provides a breakdown of news frequency on a yearly basis, it can seen that the level of news flow is highly volatile, a doubling in the level of financial news arrival which fits the filters of *economy* and *economic* is observed. This is followed by a decreases seen during the period of a fragile market recovery within the United States. This does not necessarily imply an increase in total news but only that which fits the given filters.

The levels of duplicates and reprints within news is of interest, the levels at which duplicates occur is surprising; as much as 8.8% of all news items within the corpus appear to be exact word for word duplicates, while a further 13% are identified to be near duplicates, in that they contain minor alterations. The impact of these duplicates may have a non-trivial impact on textual analysis and will be examined independently in three datasets; firstly the entire *Raw* corpus with no attempt to exclude duplicates, a second where word for word exact duplicates have been excluded, and finally a refined corpus where all articles identified as duplicates according to one of the methods outlined in section 2.4.2 have been excluded. Table 4.3 provides summary statistics for the three corpus collections for the number of individual articles per day, where No Exact Duplicates is the size of the corpus once only exact duplicates have been removed and No Duplicates where both exact and near duplicates have been removed. Even once duplicates have been removed there is a large degree of standard deviation in terms of the size of the corpus. As shown the volume of news data can vary significantly over time, in terms of the number of news items retrieved. The day with the single most unique news items of 166 took place on January 23rd 2008.

Figure 4.1 provides a graphical representation of news flow over time compressed to a

Table 4.2: Yearly Breakdown of Article Frequency.

		Duplicates		Filtered Texts	
	Raw Corpus (a)	%Duplicates (b)	%Near Duplicates (c)	(d=a-b)	(e=d-c)
2006	11,672	4.2%	13.24%	11,174	9,628
2007	20,485	11.7%	19.3%	18,087	14,145
2008	18,498	5.8%	16.6%	17,413	14,342
2009	20,210	3.5%	19.2%	19,505	15,613
2010	14,473	23.5%	4.7%	11,069	10,382
2011	11,110	8.1%	1.49%	10,210	10,044
2012	5,797	1.3%	0.7%	5,720	5,676
Total	102,245	9,067	13,348	93,178	79,830

Table 4.3: Stylized Facts of Frequency of Articles Per Day For Each Corpus.

	Raw Corpus	No Exact Duplicates	No Duplicates
Mean	42	39	33
Median	38	36	34
Standard Deviation	34	31	24
Skewness	1.2	1.2	0.7
Kurtosis	2.0	2.4	0.9
Minimum	1	1	1
Maximum	213	194	166

monthly time-frame, the significant decrease of news items within the first half of 2008 is surprising, while there is no guarantee that no errors occurred within the data curation process follow-up examination of the data flow showed no inconsistencies in the data curation process, the cause of this drop in information arrival is outside the scope of this research.

As can be seen in figure 4.2 the returns of article frequency on a monthly basis, the natural logarithm of the number of articles over the total number of articles published during the previous month can be examined. This measurement provides a means for the evaluation in rate of change at which the number of news items are released over time. There appears to be a large degree of variation in regards to the rate of news arrival, with the latter half of 2008 and early 2009 marked by a significant increase in the frequency of news items.

When examining stylized variables regarding the frequency of duplicate news items within the corpus in table 4.3 the extent to which duplicate news items occur within the corpus can be seen. On April 24^{th} 2009 as many as 116 duplicate news items were released. Of these 116 duplicates 105 were the same item being updated during the day with the changes in content being due to updated market prices. These duplicates are frequently detected through the distance algorithms since the changes from one item to the next tend to be updates of numbers. Four news items were reprinted a single time, two updated twice each and one modified three

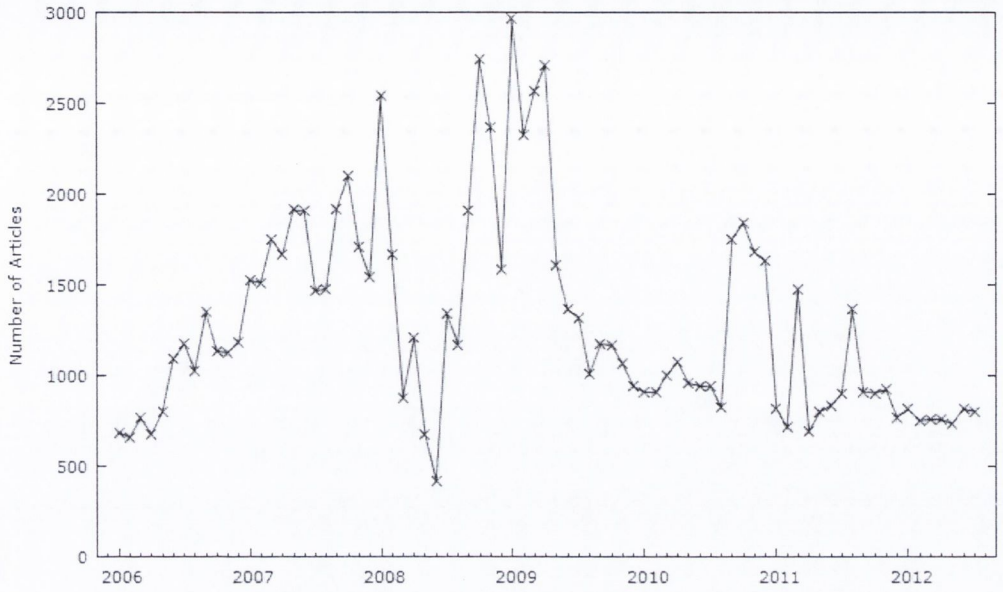


Figure 4.1: Raw Corpus Articles Per Day over Time.

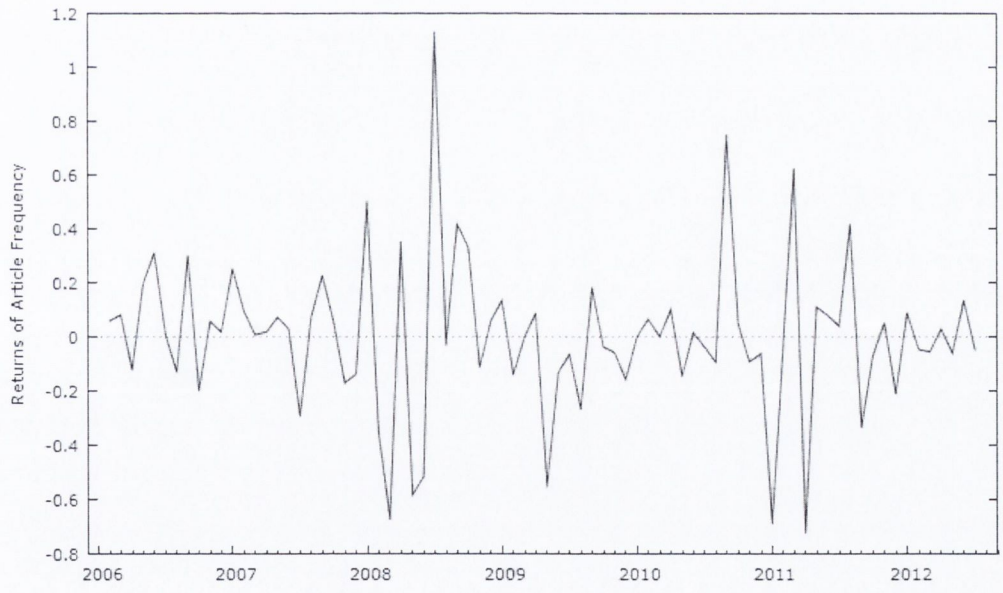


Figure 4.2: Return of Articles Per Day Compressed to Monthly.

times during the course of the day. The methods for detection of duplicates on April 24^{th} 2009 are given in table 4.4, while the distance algorithm alone is capable of finding almost all of the duplicates for time purposes the inclusion of exact equals and begins with detection methods are beneficial.

Near duplicates would suggest rather than exact reprint expansions clarifications and other edits towards previously released news items. The rapid re-releasing and re-editing of a news

Table 4.4: Detection Methods for Duplicates on April 24^{th} 2009

Method	Number of Duplicates	Percentage Detected
Exact Equals	3	2.6 %
Begins With/Ends With	2	1.7%
Distance Algorithm	105	90.5&
Longest Common Substring	6	5.2%
Total	116	

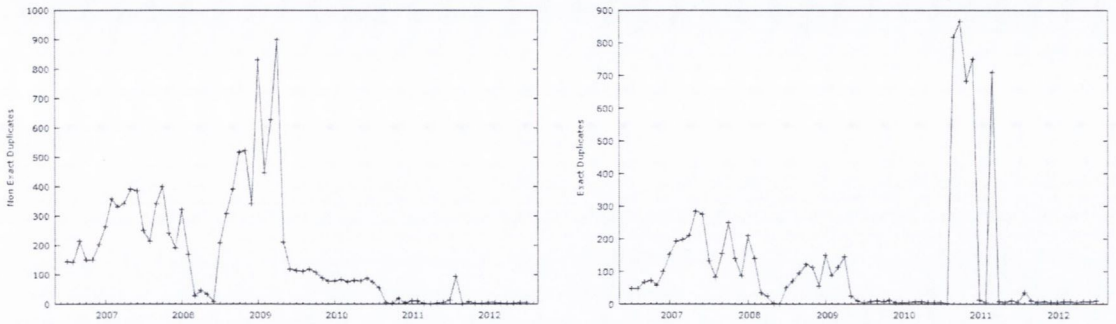
topic could suggest that the topic for discussion is of significant importance; as such these events may be of significant interest. While the repetition of an event with minor changes should present a limited amount of new information, its repetition and expansion may signify the topic's importance and thus the inclusion of such texts with the corpus may prove of interest. In regards to the reprinting of exact duplicates figure 4.3 illustrates that there was a marked increase in exact duplicates during the latter half of 2010, followed again by a rapid decrease. Table 4.5 outlines the frequency of exact duplicates and near duplicates on a daily basis, as can

Table 4.5: Stylized Facts of the Frequency of Duplicate Articles Per Day.

	Near Duplicates	Exact Duplicates
Mean	5.61	3.81
Median	1	1
Standard Deviation	10.56	8.46
Skewness	3.61	3.54
Kurtosis	17.25	13.58
Minimum	0	0
Maximum	116	59

be seen the frequency of duplicate news items per day varies greatly. The standard deviation of the number of exact duplicates is over twice that of the mean, while the frequency of near duplicate news has a ratio of 1 : 1.88.

A summary of the stylized fact of duplicate news items is presented in table 4.6, here the average number of tokens for each form of duplicate is presented. This is can be of interest as firstly is can explain if the collision detection methods are influenced by article length and if duplicates tend to be shorter or as they may be expansions of a previous news item they may indeed be longer. First the descriptive statistics of the number of tokens in individual articles is presented in table 4.6, this table is obtained from examining the total number of terms within all news items in the raw corpus with no duplicates excluded. There appears to be a significant difference in terms of the total length of news items of the corpus while the average token count is 627.84 the standard deviation is nearly 460 tokens. Clearly there is a significant degree of variation when it comes to the size of the news items in the corpus. The



(a) Non-Exact Duplicates Over Time (b) Exact Duplicates of Time

Figure 4.3: Duplicate Time-Series within NewsWire Corpus.

Table 4.6: Stylized Facts of The Number of Tokens contained within Articles in the Corpus.

	Raw Corpus
Mean	627.84
Median	586.00
Standard Deviation	459.18

next table 4.7 examines the average token frequency for each collision detected by methods. The interest here is to see firstly if the methodology appears only work in certain text sizes and if there is a difference in the types of collisions which can be seen when examining the token frequency. The table shows that for three of the four collision detection methods the size of

Table 4.7: An examination of Duplicate Article Lengths by Detection Methods

Duplicate Type	Average Duplicate Length	Difference From Average	Difference From Median
Exact Duplicates	548.8	-79	-37
Begins With/Ends With	808	175	217
Longest Substring	526	-101	-59.6
Distance Algorithm	594	-33	8

the news items is below that of the average and the mean when compared to the corpus as a whole. Only the second method *Begins With/ Ends With* returns news items much large than the average. This may likely be due to the fact that these collisions represent instances where news items have been expanded with content added to the beginning or end of a news item as shown in section2.4.2. While These results do not necessarily mean that there is some issue with these methods as the article length increases. The Exact duplicate count is below the average but since this is a simple comparison between two strings the length of the texts would have no impact on the method. Also it should be noted that from table 4.6 the standard deviation of article length within the corpus is large and these values are well below the standard deviation

by a significant factor. An examination of the nature of duplicates and the method by which they can be identified is an interesting area and worthy of future research.

4.3 Experimental Results

Presented here are the results obtained from regression analysis of the affect time series, a time series of sentiment scores obtained in the analysis, against market returns observed over the corresponding time period. The time series is generated by counting the frequency of certain affect terms as defined by one of the lexica, all news items are evaluated on a daily basis and a sentiment score is calculated. With these values ordered in a chronological manner a time series of sentiment as defined by each lexica can be created. The first area for consideration is the impact that duplicate news items within the corpus have on results obtained from the regression model. As the frequency of the near and exact duplicates is substantial their impact may have a considerable effect on the results of the analysis. The second area for examination is that of term weighting; as noted the weighting scheme applied within textual analysis may play an important role regarding the performance of a system towards textual analysis & sentiment quantification. Three schemes will be evaluated; that of a purely frequency based measurement based on absolute term frequency, and secondly that obtained using a standard term frequency inverse document frequency analysis, and a relative affect frequency where the frequency of affect terms is divided by the total number of terms.

4.3.1 Impact of Duplicates

Firstly the interest is to examine the impact, if any, that duplicate news items have on an affect time series when examined against market movements over the corresponding time period. This produces three separate time-series; one consisting of all files retried from the data-curation, a second whereby only those files identified as exact word for word duplicates are removed, and finally all those which have been identified as duplicate. In order to evaluate this time-series the regression model mentioned previously in section 3.4.2 is examined , incorporating a lag for the sentiment variable from $t-1$ to $t-5$, and also include the lag of the market returns. What follows are the results from my regression model over the time time period July of 2006 to August 2012; examining the absolute frequency of Negative terms within the General Inquirer against the Dow Jones Industrial Average using equation 4.1. The previous five data points for market returns and sentiment are regressed against next day market movements. The results from the Ordinary Least Square (OLS) regression given an estimate of the impact of each variable on market movements. Since the interest is in evaluating a statistical relationship between sentiment and market movements results are examined in a manner similar to Tetlock.

Two variables of interest are the coefficient of determination (R^2) from the regression models and the P-value for sentiment. R^2 gives a statistical measure of how well the regression line approximates the real data points. $P-Value(Sent)$ gives the significance of the sentiment values in the regression model, estimating if their inclusion shows a statistically signification relationship.

$$Dow_t = \alpha 1 + \beta 1.L5(Dow_t) + \gamma 1.L5(Sent_t) \tag{4.1}$$

Table 4.8: Impact of Duplicates When Regressing Against the Dow Jones Industrial Average. This table presents the Ordinary Least Square (OLS) regression results of the Dow Jones Industrial Average (DJIA) across negative sentiment extracted from our AP financial Newswire corpora examining the impact on the inclusion and exclusion of duplicate news items. Negative sentiment is measured as an absolute term frequency count of terms as included in the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% confidence level(***).

Sentiment Measurement	No Near Duplicates	No Exact Duplicates	Raw Corpus
Sent t_{-1}	-13.4	***-29.1******	***-28******
Sent t_{-2}	9.9	13.5	13.7
Sent t_{-3}	-1.8	5.9	6.6
Sent t_{-4}	9.7	8.5	7.3
Sent t_{-5}	*-12.4**	-6.1	-5.7
Dow t_{-1}	***-18.8******	***-19.3******	***-19******
Dow t_{-2}	-11.3	-11.3	-11.2
Dow t_{-3}	4.6	5.3	5.7
Dow t_{-4}	-2.4	-2.1	-2.1
Dow t_{-5}	-9.2	-9.3	-9.2
R^2	3.53%	4.57%	4.55%
AIC	-8700	-8716.5	-8716.2
$P-Value(Sent)$	0.32	**0.05****	**0.03****

An examination of the results identifies two issues, firstly that while the sentiment at time $t-1$ is significant to the 99% within two of the proposed regression models while the third model having excluded all duplicates is not statistically significant, values beyond $t-1$ are statistically insignificant, while the reversal of sign from negative to positive is the same as observed by Tetlock [104] with the level of statistical significance dropping beyond $t-1$. Such a result indicates that an increase of negative sentiment is followed by a downward movement within markets on the following day, such as would be expected and has been observed in previous works. Tetlock argues that the reversal of the sign from negative to positive indicates that the impact of sentiment is short lived and beyond day before there is a return to market fundamental, in both his work and the results here there is a reversal of the sign of the coefficient

but in both cases the impact is not as statistically significant. Tetlock argues that while there is a return to market fundamentals the reversal is not significant enough to offset the initial decline.

There is a strong statistical significance (to 99% confidence level) between market returns at time t and past market returns but only to a lag of $t-1$. Results are presented in a manner similar to Tetlock where the coefficients of regression indicates the average impact of a one-standard deviation change in negative sentiment as defined by the General Inquirer.

Secondly there is a significant difference regarding the performance of the dataset which includes duplicates when compared to that without. The $p-value$ for the null hypothesis that the five lags of negative sentiment does not forecast returns ($P-Value(Sent)$) is 0.05 and 0.03 for the corpus excluding exact duplicates and including all duplicates respectively which implies that negative sentiment may be related to future market returns. The exclusion of exact duplicates increases the coefficient at time $t-1$ by one basis point. Indeed while the datasets with some form of duplicates are statistically significant to 95%, there is no significance in the refined dataset with all duplicates excluded.

At this stage it is useful to examine the relative performance of one model when compared with another, while the R^2 values do show which performs in a superior way, aiming to quantify the level to which one performs better than alternatives. An examination of the Akaike Information Criterion (AIC) allows for an evaluate the performance of one model when compared to another. It provides the relative measurement of the performance of a statistical model when compared to other models for a given data set. While this will in itself not be able to provide us with an indication regarding exactly how well the model performs it does allow for a means of selecting the *best* performing among the proposed models.

The AIC value may be calculated according to formula 4.2:

$$AIC = 2k - 2ln(L) \tag{4.2}$$

Where k is the number of parameters and L is the maximized value of the likelihood function of the regression model. In inclusion of k means that the AIC value of a model is not based only on how well it fits the data, but takes into account the number of variables included. The penalty applied when a model increases the number of parameter k is to account for overfitting. Once the AIC has been calculated for each model each may be compared.Firstly the difference between each model Δ_i is calculated and the lowest AIC; that obtained from the *best* model calculating where $\Delta_i = AIC_i - minAIC$. While this provides a relative value when compared to the other given model a superior method for their interpretation is to normalise the relative values as weights; *Akaike weights* w_i.

The *Akaike weights* may be calculated as:

$$w_i = \frac{exp(-0.5 * \Delta_i)}{\sum_{r=1}^{R} exp(-0.5 * \Delta_r)} \tag{4.3}$$

Table 4.9: DJIA AIC Weight Calculation.
We calculate the AIC weights to evaluate the effect of the exclusion of duplicates and near duplicates within our corpus. These results are obtained from an examination of absolute negative affect frequency within the three AP financial Newswire corpora, with negative terms identified according to the GI regressed against the DJIA. Results in italics indicate that the relative weight of the model indicates that the model can be excluded from further analysis.

Model	*AIC*	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
No Exact Duplicates	-8716.52	0	1	0.53
Raw Corpus	-8716.21	0.30	0.947	0.46
No Near Duplicates	*-8700.07*	*16.45*	*2E-4*	*0*

The relative weights of the results obtained from each of the three models and datasets are presented in table 4.9, these results demonstrate to what extent the exclusion of all duplicate news items have on the regression modeling. The value W_i is the probability that model i provides the best approximation, clearly here the model with all duplicates excluded is significantly poorer in performance. Commonly when evaluating the weights of multiple models those with a value W_i which is not within 10% of the highest may be excluded [87]. While the values obtained from the corpus with no exact duplicates is slightly superior there is no evidence to conclusively state that the inclusion of exact duplicates has a negative impact in the affect time-series generation.

The above experiments are expanded to examine if the observed results apply to alternative market indices. In the interest of brevity the analytical results are presented in appendix B. When examining the role of sentiment against the S&P 500 there is a statistical significance link between future market returns at time $t-1$ (with a significance level of 99%) and negative affect as defined by the General Inquirer (table B.1). This significance however is seen only against the affect time series obtained from the corpus allowing for the inclusion of duplicate news items (AIC weights presented in table B.2). Again similar to the DJIA analysis the coefficient at time $t-1$ is negative which is followed by a reversal to positive, this suggests, as discussed by Tetlock that the impact of negative sentiment is short lived and followed by a return to market fundamentals. Indeed an increase of one standard deviation in sentiment at time $t-1$ has an impact of 30 basis points on the following days market returns when examining the corpus comprised of no exact duplicate news items. The inclusion of sentiment within the regression model is significant to the 95% confidence level when some form of duplicates are included but is statistically insignificant once duplicates are excluded.

When examined against the VIX index (table B.3), this index is of interest, as it has been examined in previous works within sentiment analysis demonstrating a statistically significant link with a negative affect time series produced through an examination of Google search queries [17]. This series shows a much lower level of statistical significance, where the inclusion of sentiment lags is significant to 90% only when examining the corpus containing no exact duplicates, however sentiment at time $t-1$ is significant to 95% for corpora containing some form of duplicates. A striking contrast between the analysis of the VIX and previous market indices is that while the DJIA and S&P 500 showed little statistical significance with previous market returns beyond time $t-1$ the VIX demonstrates a level of statistical significance for all lagged return values. A further contrast is that while both previous results have displayed a negative coefficient at time $t-1$ indicating a fall in returns following an increase in negative sentiment, within the VIX this sign is positive followed by a reversal to negative at times beyond $t-1$. Such a result indicates that a increase in negative affect will result in an increase in VIX returns. As the VIX is a measurement of market volatility witnessed within the S& P 500 such a result suggest that negative sentiment has an impact on investors' fears and as such volatility increases, as has been witnessed in previous studies examining the role of sentiment on market returns and volatility [5]. However as the level of significance is somewhat weak within these results it may not necessarily draw any definitive conclusions (AIC weights calculations presented in table B.4). It has been demonstrated that negative affect as defined from the GI extracted from two of the three corpora has a statistically impact on next day market returns, the effect is followed by a reversal of the sign of the affect coefficient and the level of significance drops sharply beyond one day. Across the three market indices the inclusion of duplicate news items within the corpus for examination has a statistically significant impact on the analysis, and while there is some improvement through the exclusion of identical news items where an news article has been replicated in-toto the difference is not sufficient to draw lasting conclusions. Unless otherwise indicated all future results are those obtained from an examination of the corpus containing no exact duplicates.

4.3.2 Impact of Weighting

The impact of an alternative weighting metrics applied to the affect time-series is examined. As the previous results have outlined there is a statistically significant impact on the inclusion of duplicates within the corpus for sentiment analysisan examination of three weighting function is applied for the extraction of sentiment from the news corpus which include no exact duplicates. A simple feature frequency (absolute) term count is applied, a $t_f - i_{df}$ weighting function and a relative term frequency analysis. Within relative term frequency all articles occurring on a single day are summed and examined as a single large news item. Relative

term frequency was included for two reason, earlier in this research I examined the impact of relative term frequency extracted from Irish financial news and its impact on the Irish financial markets [18]. This measurement was also included as it was used in Tetlock's research in the area of sentiment analysis, however in Tetlock's research where he examined a single news item each day (Abreast Of The Market) while here the number of news items returned from the search criteria varies on a day to day basis. While relative term frequency evaluates news items as a single text document summing the length $Tf - Idf$ score calculation factor in the number of news items. This measurement was included as the usage of such a weighting function is common within textual analysis and was also applied in the research by Loughran and McDonald [64].

Table 4.10: Impact of Sentiment Score Computation Against DJIA.

This table presents the OLS regression results of the DJIA across negative sentiment extracted from our AP financial Newswire corpus containing no exact Duplicates. We compare the impact of three weighting functions Absolute negative term frequency, relative term frequency and $tf - idf$ as defined by the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Absolute	Tf-Idf	Relative
Sent t_{-1}	***-29.1******	***-27.9******	*-8.1**
Sent t_{-2}	13.5	13.4	5.07
Sent t_{-3}	5.9	4.8	-1.4
Sent t_{-4}	8.5	8.5	6.7
Sent t_{-5}	-6.2	-6.9	-6.5
Dow t_{-1}	***-19.3******	***-19.3******	***-18.9******
Dow t_{-2}	-11.3	-11.3	-11.2
Dow t_{-3}	5.4	5.2	4.9
Dow t_{-4}	-2.2	-2.3	-1.6
Dow t_{-5}	-9.4	-9.7	-8.5
R^2	4.57%	4.52%	3.17%
AIC	-8716.5	-8715.8	-8694.3
$P - Value(Sent)$	**0.05****	**0.04****	0.17

An examination of these three weighting functions against the DJIA is presented in table 4.10 Across all three weighting function there appears to be little to no statistical significance beyond time $t - 1$. While there appears to be little difference between the results obtained through absolute term frequency and $t_f - i_{df}$ there is a significant difference when examining the relative term frequency. While sentiment at time $t - 1$ is significance to the 99% across absolute term frequency and relative term frequency within relative term frequency this is only significant to 90% level. And while the inclusion of sentiment is significance to 95% for both

absolute and tf-idf, the $p - value$ for relative frequency is 0.17 as presented in table 4.10.

When examining the AIC weights in table 4.11 it can be seen that the relative weighting method may be excluded, however while absolute term frequency is superior to $tf - idf$ it is not to a significant degree. Similarly the R^2 values while better in the absolute term frequency time series (4.57%) when compared to $tf - idf$ (4.52%) it is not sufficient to draw a lasting conclusion regarding the impact of these two weighting methods.

Table 4.11: AIC Sentiment Score Computation Analysis for DJIA and Negative Affect
We calculate the AIC weights to evaluate the performance of each of the three weighting methods absolute term frequency, tf-idf and relative term frequency of negative affect terms as defined by the GI regressed against the DJIA, term frequency is obtained from our corpus of no exact duplicates from the retrieved AP Financial Newswire. Results in italics may be excluded from further analysis.

Weighting Method	*AIC*	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
Absolute	-8716.52	0.00	1.00	0.58
Tf-Idf	-8715.84	0.68	0.71	0.42
Relative	*-8694.32*	*22.2*	*0.00*	*0.00*

Similar results are obtained when examining the S&P 500 and VIX (presented in tables A.1 & A.3), for the S&P 500 the inclusion of sentiment is significant to 95% for both absolute term frequency and $tf - idf$ the $p - value$ for the relative affect time series is 0.15. The significance level observed against the VIX is to a lower extent (90% for the absolute term frequency) there is a improvement over the other weighting methods. When examining the AIC weights (tables A.2 & A.4) relative affect frequency can be excluded yet find similar results between absolute frequency and tf-idf.

Initial Findings and Discussion

The above findings suggest a statistically significant relationship between negative affect sentiment and future returns in each of the three given indices. In each instance the inclusion of duplicates appear to have a significant impact on the analysis of the resulting affect time-series, while there appears to be some improvement from the exclusion of identical news items shown in tables 4.8, B.1 & B.3 the improvement is not sufficiently significant to draw any major conclusions. The cause of this improvement and the apparent impact of duplicates is interesting and while outside the scope of this research is certainly an area for future consideration. There is the potential that not only does the reprinting and expansion of a news item have in itself some impact, but that news sources will expand and recommend on a topic or issue purely if it is of importance, however such comments are at present purely speculation.

The coefficient of negative affect at time $t - 1$ across the three given indices is illustrated in figure 4.4 in basis points, illustrating the impact of a change of one standard deviation on

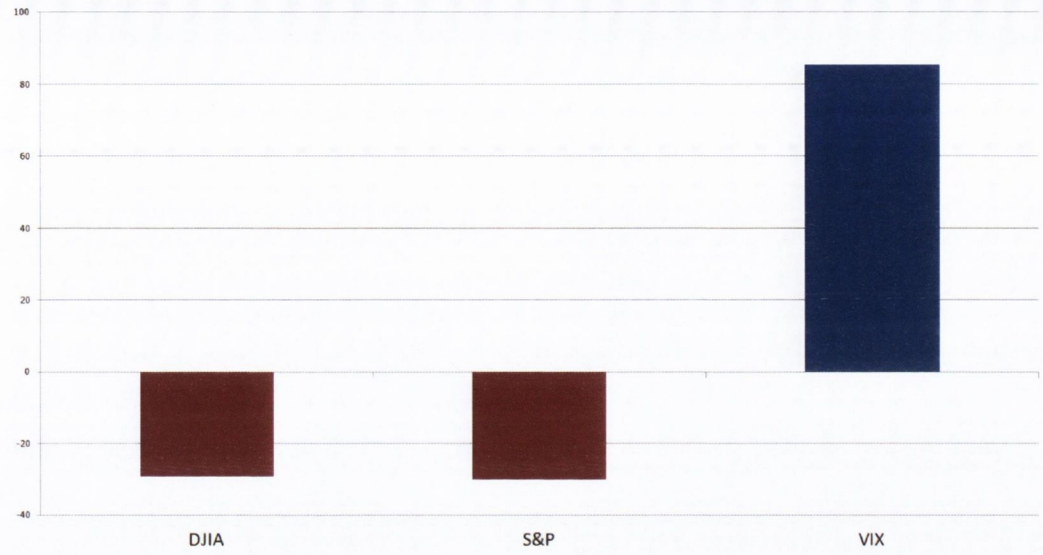


Figure 4.4: $T-1$ Coefficients for Negative Affect at $t-1$

The coefficient γ_1 from the given regression model of absolute negative affect frequency contained within our corpus of no exact duplicates regressed against the given index. Each coefficient measures the impact of an increase of one standard deviation of negative affect on returns in basis points (one basis point equals a return of 0.01%). Results are obtained from Absolute Negative term frequency from our corpus with no exact duplicates.

market returns on the following day. Within both the DJIA and S&P 500 an increase in negative affect will cause a downward movement on market returns on the following day, while for the VIX index this increase will results in an increase in returns. As the VIX is a measurement on volatility within the S&P 500 this suggests that negative affect results in an increase of volatility within this index.

While the impact of sentiment on financial indices is modest these findings do propose a relationship between the two which have been noted in previous research yet there is at times a temptation within current literature to overstate the potential value of such information. Indeed while the resulting R^2 values are low they remain statistically significant.

And while some have proposed the development of trading systems based on sentiment [9, 90] and have at times shown encouraging results, other have demonstrated that a system to determine market trading strategies based on sentiment extracted is flawed and unlikely to consistently produce positive returns over time [51].

In the interest of brevity results that following will concentrate on sentiment produced using absolute term frequency extracted from the corpus containing no exact duplicate unless otherwise stated due to the fact that the results obtained through this corpus and weighting method appear to produce slightly superior results over tf-idf and the inclusion of exact duplicates within the collected corpus. Also as the area of interest is an examination of the sentiment coefficients γ_n the coefficients for market returns α will be excluded from future

tables.

4.3.3 Alternative Lexica

Having considered the impact of weighting schemes and the role of duplicates and near duplicates within this corpus the evaluation is expanded to examine the performance of alternative lexica within sentiment analysis of financial news against the aforementioned market indices.

Initially the regression model used within the previous evaluations including lagged values of sentiment and market returns to a time $t-5$ will be used. Based on the previous finding regarding the role of duplicates; in that the inclusion of duplicates has a statistically significant impact on regression analysis all further results will be those obtained through and examination of our corpus which contain no exact duplicate news items, while there is a modest improvement through the exclusion of non-exact duplicates it is not however sufficiently significant to draw lasting conclusions regarding their impact. Similarly as there appears to be little difference in results obtained using absolute term frequency and tf-idf, unless indicated all future results presented will be those obtained using absolute term frequency with no weighting function applied.

The lexica selected for examination offer a number of issues to be examined, firstly the expansion of the negative affect lexicon to include stemmed version of all terms, while the impact of stemming term may be viewed as superior as it will identify terms which while having similar meanings may not be found within the initial lexica and are the result purely of grammatical, rules this improvement is by no means guaranteed [59, 69], furthermore this lexica is include as the larger size of it posses extra challenges when implementing methods to refine this lexica. As the feature size increases (in this instance the number of individual terms) optimization techniques such as genetic algorithms and particle swarm optimization may run into issues whereby the large feature space makes the convergence of the system less likely as discussed in section 2.5.1. The choice of lexica also allow an examination of positive affect frequency; this examines the previously observed results where positive affect frequency while having an impact on future market returns is shown to be less significant when compared to negative affect frequency. Lastly is an examination of an alternative lexica developed within a domain specific context, in this instance financial reports. This allows for an evaluation of legacy lexica developed for general language against domain specific lexica, intuitively it may be assume that a lexica developed through expert knowledge of language and the domain would capture sentiment to a greater extent, as term misclassification is less likely while terms which may not be included within a general language lexica due to their domain specific nature are included. However as noted within other areas of textual and sentiment analysis

such suppositions are not always found to be true. While an examination of the language within a corpus does provide insight when developing a domain specific lexica through human intervention, on occasion terms with the greatest information have not been included within those put forward by human respondents. Firstly the performance of varying lexica against the DJIA is examined, where previously the greatest level of statistical significance when compared to results obtained from the other proposed indices was found.

Table 4.12: Regression of Varying Dictionaries against DJIA.

The table shows the Ordinary Least Squared (OLS) regression results of the Dow Jones Industrial Average (DJIA) across the four dictionaries. The affect time series is extracted through an absolute term frequency count. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%), the regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) while bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	***-29.2******	***-29.2*****	***-28.9*****	***-25.8*****
Sent t_{-2}	13.5	14.8	8.4	11.1
Sent t_{-3}	6.0	5.8	11.8	5.0
Sent t_{-4}	8.5	6.1	5.5	2.6
Sent t_{-5}	-6.2	-4.7	-3.1	-0.9
R^2	4.57%	4.53%	4.24%	4.14%
AIC	-8716	-8715	-8711	-8709
$P-Value(Sent)$	**0.05****	*0.06**	0.138	0.13

Similar to previous examination the impact of the sentiment score on the day before is significant (to the 99% confidence level), however at time lags beyond this there does not appear to be a statistically significant impact on the DJIA. In each instance the coefficient of the initially lagged value $t = -1$ is negative while those following are positive; again alluding to the return to market fundamentals. This holds true even when examining the frequency of positive terms within the given corpus.

While researchers have stated that positive words contain less information than negatives [31] the negative sign of the *positive affect* coefficient suggests that an increase in positive affect term frequency is followed by a downward market movement on the following day. The results are counter intuitive in that the sign of the regression coefficient is negative which would suggest that an increase in positive affect term frequency results in a downward movement in markets. The intuitive judgment would suggest that the positive sentiment should have a positive coefficient and result in an upward movement if there is a relationship between financial markets and the sentiment measurement. While the coefficients at $t-1$ is statistically significant the impact of the sentiment time series is not with a p-value of 0.138 against the DJIA. This is in contrast to the Negative affect score and the Negative-Inf score which

Table 4.13: Regression of Varying Dictionaries against S&P 500.

The table shows the Ordinary Least Squared (OLS) regression results of the Standard & Poor's 500 (S&P 500) across the four dictionaries. The affect time series is extracted through an absolute term frequency count. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%), the regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) while bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	***-30.01******	***-30.36******	**-29.36****	**-27.46*****
Sent t_{-2}	12.91	14.30	7.02	9.83
Sent t_{-3}	4.13	4.01	11.01	3.52
Sent t_{-4}	11.05	8.32	8.09	4.56
Sent t_{-5}	-6.87	-5.12	-4.42	-0.21
R^2	4.27%	4.24%	4.0%	3.9%
AIC	-8427.79	-8427.253	-8422.82	-8422.41
$P-Value(Sent)$	**0.05****	*0.061**	0.15	0.14

are significant to 95% and 90% respectively. While the results are counter intuitive given the P-Value for positive sentiment it may not be stated that there is a statistically significant relationship between my positive affect timeseries approach and future market movements.

The performance of the Financial Negative (*Fin Neg*) dictionary is striking; such a reference lexicon should in theory capture financial reporting to a greater degree having taken into account the specific meaning of terms when applied towards the financial domain. However it should be noted that the authors did not suggest that such this lexicon accurately reflects general financial news language but rather that contained within a more technical report format. However it would be intuitive to assume that terminology accurately defined within technical financial reporting is used within a different context when found within financial news reporting. Yet there is the potential that the terms found within the financial negative dictionary are of a technical nature which would not be found within general language reporting on economic affairs, and as such this would lower the volume of incoming affect terms.

A comparison of the R^2 values across the four given reference lexica identifies the greatest values obtained from the negative affect lexicon which is slightly superior to a stemmed version of the negative lexicon.

The regression results against the S&P 500 across the given lexica presented in table 4.13 and are in-line with those previously observed. Again there is a greater level of statistical significance when comparing negative affect than that obtained from positive affect term frequency. While again the financial negative lexicon produces modest results, the impact of weighting appears to have little impact on the resulting analysis.

Again the analysis against the VIX performs poorly when compared to the previous indices;

Table 4.14: Regression of Varying Dictionaries against VIX.
The table shows the Ordinary Least Squared (OLS) regression results of the Chicago Board Options Exchange Market Volatility Index (VIX) across the four dictionaries. The affect time series is extracted through an absolute term frequency count. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%), the regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) while bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	**85.3****	**87.6****	*75.6**	**88.2*****
Sent t_{-2}	-37.5	-37.5	-16	-26.5
Sent t_{-3}	22.25	6.1	-3.9	-5.7
Sent t_{-4}	-45.2	-45.9	-43.6	-31.8
Sent t_{-5}	3.3	-1.7	-1.7	-17.1
R^2	3.68%	3.71%	3.50%	3.70%
AIC	-3719.96	-3720.53	-3717.23	-3720.27
$P-Value(Sent)$	0.11	*0.09**	0.4	*0.084**

with some level of statistical significance with the negative lexica yet little evidence against alternative given lexica (table 4.14). Across each of the lexica there continues to be a reversal of signs of the coefficients when compared to those see against the DJIA and S&P 500 regression models (tables 4.12 and 4.13). Again indicating that an increase in frequency across any of the given lexica results in an increase in future returns yet is again reversed later in the week where the coefficient turns negative, this is in contrast to the two alternative indices which are followed by a downward motion on the following day. The regression of sentiment against the VIX appears to confirm that it is indeed such an index if only because the regression co-efficient for the day before is positive across the four lexica used individually (GI Negative, Negative-Inf, GI Positive and Financial Negative). The key difference here is that the statistical significance is not the same as my previous results for the DJIA (table 4.12) and S&P500 (table 4.13). In that first the regression co-efficient is only significant to the 90% confidence level and second this is only achieved when using the extended term list by Loughran & McDonald or their financial negative term list.

4.3.4 Decreased Lag Of Sentiment

As there appears to be no statistical significance in any of the regression series beyond the impact of the day before sentiment it may be of interest to examine the impact of simplifying the model by reducing the window of lagged sentiment values. Below is an examination of the impact of a shorter lag across all the dictionaries, these models are simplified whereby only the values for the previous two days are included within the regression model. Two days are

Table 4.15: Shorter Time Lag Regression of Varying Dictionaries against DJIA.

The table shows the OLS regression results of the DJIA against absolute frequency extracted from the four dictionaries of AP financial newswire articles with no exact duplicates included reducing the previously given regression model to include sentiment and market returns to a lag of 2. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%), The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) while bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Negative	Negative-Inf	Positive	Fin Neg
$Sent_{t-1}$	***-26.82******	***-26.66*****	***-23.46****	***-23.03*****
$Sent_{t-2}$	**18.28****	**18.29****	*15.62**	*13.84**
R^2	3.72%	3.74%	3.28%	3.45%
AIC	-8734.3	-8734.6	-8727.29	-8730.1
$P-Value(Sent)$	***0.008*****	***0.01*****	*0.07**	**0.025****

picked so as to examine if there is an apparent reversal of the impact of sentiment on market movements, Tetlock identifies that at a lag beyond $t = -1$ there appears to be a reversal in the coefficient of sentiment stated to be a result to market fundamentals [104].

Table 4.15 outlines the results of the new simplified regression model, when compared to the results within table 4.8 an improvement in with a significance increase from 95% to 99% across the first two datasets is seen. With the lagged value of $t = -2$ being significant to 95% suggesting a potential impact of sentiment beyond $t = -1$ while again witnessing the reversal from a negative coefficient to positive. Note that the results reported in section 4.3.3 and 4.3.4 are based on the use of a corpus where exact duplicates have been removed. The impact of keeping or removing duplicates on the regression coefficients and by implication the fitness function is small (The results of these difference corpora types is in appendix B). These results are continued across the alternatively proposed lexica seen in table 4.15 indeed there is a significant improvement seen for both the positive affect time series results and the negative finance time-series.

Firstly that all are statistically significant to 95% and above, and as has been shown in previous studies the results from negative affect appear to be more statistically significant that those obtained using positive affect frequency. Of interest is that there appears to be little difference between the results obtained though a stemmed version of the negative dictionary and those obtained through the un-stemmed raw negative term list obtained from the General Inquirer. Secondly the Negative Finance dictionary provided from Loughran & McDonald does not perform as well, while this dictionary was proposed for financial reports it may be assumed that the correct classification of terms when applied to more complex financial reports should hold true when applied to financial news reporting. In the production of this lexicon specifically

tailored to financial terms approximately 73% of terms in the GI Negative category were removed as they were viewed to be incorrectly classified when it came to financial language.

Note positive result is only significant to 90% confidence level where as the impact of negative sentiment is at 99% confidence level and financial negative to 95%. The results for the inclusion of lagged values of sentiment beyond $t = -1$ for negative affect are compared as a means for examining if it may be concluded that lags beyond $t - 2$ may be excluded for the purpose of future analysis.

Table 4.16: AIC Analysis of Shorter Lag for Negative Affect Against DJIA
The table below evaluates the performance of the examination of negative affect against the DJIA where the time tag of the past sentiment and market returns is reduced from $t - 5$ to $t - 2$. Values in italics may be excluded from further examination.

Maximum Lag	*AIC*	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
Lag $t-2$	-8734.3	0	1	0.99
Lag $t-5$	*-8716*	*18.3*	*0*	*0*

Through an examination of AIC weights in table 4.16 the impact of the reduction of the maximum time lag from $t - 5$ to $t - 2$ can be compared, the results indicate that all values beyond $t - 2$ may be excluded from future analysis as there appears to be little to no little significance.

The inclusion of $t - 2$ remains statistically significant and through the sign reversal maintains the indication that the impact of sentiment is reversed after its initial impact. These results are similarly seen when examining negative affect against both the S&P 500 and VIX (tables E.5 and E.9). As such within all future analysis the maximum time is limited to a lag to $t - 2$.

4.4 Evaluation

Following is an evaluation of the implementation of term feature selection from each of the proposed lexica with the aim of reducing the number of affect terms across increased lexica with the aim of eliminating terms which within this domain are misclassified and its potential for the improvement of regression analysis.

The previous section presented an examination of lexica and the impact of weighting lagged values and duplicates within the corpus against the three proposed indices. These results obtained from *Raw* lexica provide a base-line for an examination of the performance increase through the usage of refined lexica as produced through biologically inspired algorithms. This section will concentrate purely on results obtained using a corpus containing no exact duplicates as the previous results show this corpus tends to produce the optimal results, and will also concentrate on results obtained using absolute frequency counts due to the observation that

relative frequency performs poorly while there is a small improvement in the implementation of absolute frequency count when compared to tf-idf. Due to the previously observed superior results obtained from the reducing of maximum lag to $t-2$ all further results will concentrate on the modified regression model. However the training method used for the development of refined lexica used the original model as an fitness function with the obtained R^2 applied for fitness value. However the analysis from the original time frame is presented in appendix D.

4.4.1 Refined Lexicon Performance

The first lexicon for examination is the negative affect category as found in the General Inquirer. The regression results from the original dictionary are compared with the results from the two refined dictionaries. The complete regression analysis results are presented in appendix E in tables E.1, E.5 and E.9. Following is a comparison of the performance of the *Raw* negative lexicon obtained from the General Inquirer and those obtained through feature selection using both genetic algorithms and particle swarm optimization in table 4.17.

Table 4.17: Comparison of Refined Negative Lexicon DJIA

This table presents the OLS regression results of the DJIA across negative sentiment extracted from the AP financial Newswire. Negative sentiment is produced from a raw absolute frequency count of terms as defined against the General Inquirer and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations. I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Sentiment Measurement	Negative	Negative $_{Genetic}$	Negative $_{Particle}$
Sent t_{-1}	***-26.8****	***-28.96****	***-29.51****
Sent t_{-2}	**18.28****	**18.65****	**20.73****
R^2	3.72%	4.22%	4.06%
AIC	-8734.282	-8742.251	-8739.632
$P-Value(Sent)$	***0.008****	***0.006****	***0.003****

Both the lexica developed through genetic algorithms and particle swarm appear to perform superior to the *Raw* General Inquirer negative lexicon. The signs of the coefficients remain the same; a negative at $t-1$ indicating a downward pressure on market returns followed by a reversal. When comparing the predictive value of the regression models an increase of in the R^2 value of 13% for genetics (3.72% increasing to 4.22%) and 8% for particle swarm (3.72% increasing to 4.06%) can be seen. By examining the AIC weights obtained from the three models it may concluded that the negative GI lexicon can be excluded with a weight of 0.005, and while there is a significant improvement seen within the Negative $_{Genetic}$ word list it is not sufficiently superior to that obtained from particle swarm.

This is followed with an examination of the performance of negative affect and feature selection against the S&P 500 index in table 4.18.

Table 4.18: Comparison of Refined Negative Lexicon against S&P 500

This table presents the OLS regression results of the S&P 500 across negative sentiment extracted from the AP financial Newswire. Negative sentiment is produced from a raw absolute frequency count of terms as defined against the General Inquirer and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Sentiment Measurement	Negative	Negative $_{Genetic}$	Negative $_{Particle}$
Sent t_{-1}	**-27.91*** **	**-30.78*** **	**-30.53*** **
Sent t_{-1}	**17.92** **	**19.97** **	**20.14** **
R^2	3.64%	4.20%	3.95%
AIC	-8449.096	-8457.995	-8454.008
$P-Value(Sent)$	**0.012** **	**0.002*** **	**0.004*** **

Again there is an improvement in results when examining returns in the S&P 500, the obtained $P-Value$ for the impact of a lagged value of sentiment increases from 95% confidence within the raw lexicon to 99% confidence within both obtained through feature selection. Which leads to an improvement in the R^2 values of 15% and 7.4% for genetics and particle swarm repetitively (Full analysis in table E.5). As with the DJIA when comparing the AIC weights obtained from the models the raw negative lexicon (with a weight of 0.01) may be excluded while again the improvement obtained through genetic algorithms is superior it is not to the extent that particle swarm (weights 0.87 and 0.12 respectively) can be excluded. From this it can seen how the refinement process has indeed improved a number of the regression analysis results when using the refined negative affect lexicon in particular the results of the genetic algorithm based refinement.

There does not appear to be a significant improvement when examining the VIX index, indeed the implementation of a refined lexicon has a detrimental impact on the regression analysis. While there is little evidence of a statistically significant relationship between my sentiment based time series and the VIX as shown in table 4.19 whereby the affect time series obtained from the GI is significant to 90% (full analysis in table E.9). This is not a definitive result that there is no relationship, simply that none was found using the machine learning techniques and lexica in my implementation. If there was no statistical relationship in the method proposed it is possible that the techniques applied attempted to overfit the data. It is possible that as the VIX is a somewhat more complex measurement of financial indices measuring volatility in the S&P 500. By attempting to force a solution in my approach the

Table 4.19: Comparison of Refined Negative Lexicon against VIX

This table presents the OLS regression results of the VIX across negative sentiment extracted from the AP financial Newswire. Negative sentiment is produced from a raw absolute frequency count of terms as defined against the General Inquirer and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Sentiment Measurement	Negative	Negative $_{Genetic}$	Negative $_{Particle}$
Sent t_{-1}	**73.09****	**60.41****	**71.02****
Sent t_{-2}	**-61.73****	**-48.12****	*-58.02**
R^2	2.52%	2.42%	2.48%
AIC	-3723.6	-3722.1	-3723.1
$P-Value(Sent)$	*0.08**	*0.07**	0.11

evolved solution does not describe any underlying relationship and as such would perform poorly against data not used in its training. This is illustrated graphically in figure 4.5, where the relative change in R^2 across the three given indices is shown when the proposed feature selection methods are applied.

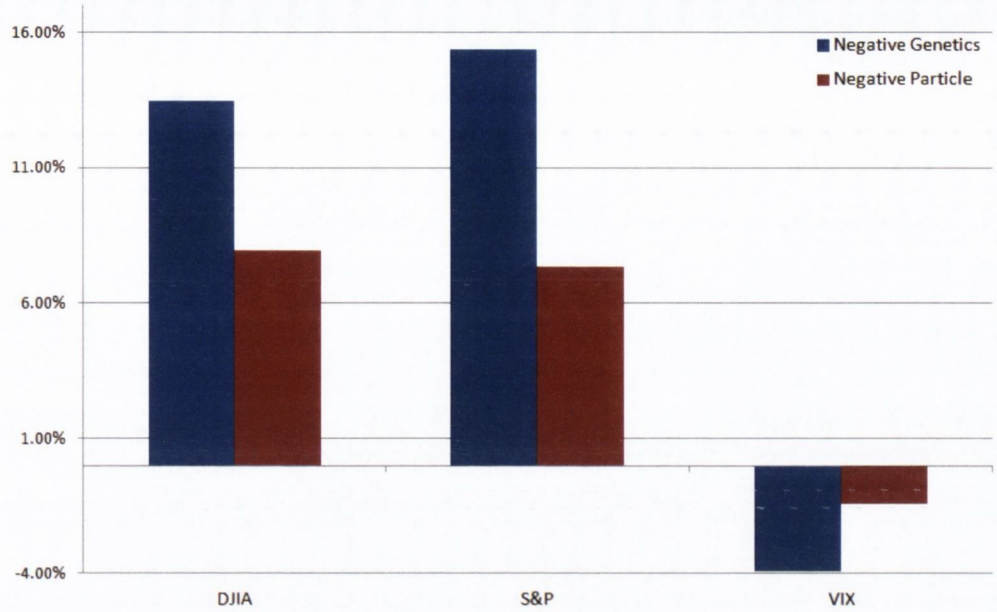
Following is an examination on the impact of feature selection applied to the stemmed version of the negative affect lexicon. As previously discussed this lexicon produces issues in that the feature space is large with a total of 4,187 individual terms which as discussed may prove problematic towards both genetic algorithms and particle swarm.

Table 4.20: Comparison of Refined Negative-Inf against DJIA

This table presents the OLS regression results of the DJIA across negative sentiment extracted from the AP financial Newswire. Negative sentiment is produced from a raw absolute frequency count of terms as defined against the stemmed version of the General Inquirer and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Sentiment Measurement	Negative-Inf	Negative-Inf $_{Genetic}$	Negative-Inf $_{Particle}$
$Sent_{t-1}$	***-26.66****	***-28.50****	***-28.75****
$Sent_{t-2}$	**18.29****	**19.91****	**20.37****
R^2	3.74%	4.2%	4.00%
AIC	-8734.591	-8741.993	-8738.667
$P-Value(Sent)$	***0.01****	***0.002****	***0.01****

Table 4.20 illustrates that the refined lexica have a superior performance to that obtained through the raw GI lexicon, in this instance there is a 12% improvement in the R^2 value



Here I illustrate the impact of the feature selection methods on the predictive R^2 values obtained from regression analysis of absolute negative affect frequency within the corpus of no exact duplicates, against the three given indices.

Figure 4.5: R^2 improvement in Negative Affect regression across each index with both feature selection methods

when implementing genetic algorithms and a 6% improvement when implementing particle swarm. These results indicate that while the feature space does contain a significant number of variables the proposed methodology appears to have successfully *evolved* a solution which produces superior results within this instance. However this is not always the case where the genetic algorithm approach to the negative+ dictionary applied to the S&P 500 has a (minor) detrimental impact, reducing the R^2 value by 2% (table E.6), however the particle swarm successfully produced a superior lexicon with R^2 value of 3.94% an improvement of 7%. A similar outcome is observed against the VIX (tableE.10) where the lexicon produced through genetic algorithms is somewhat inferior the raw GI dictionary while the lexicon produced through particle swarm is somewhat superior (to the degree of-2.5% and 3.24% respectively). However there remains little indication of statistical significance between the affect time series and future movements within the VIX index.

Following is an examination of the performance of the positive affect time series regressed against the DJIA, as has been noted positive affect is deemed to have a lesser impact on investor sentiment and as such any analysis is proposed to produce results which predict market returns to a lesser extent. An examination of the performance of positive affect and the refined lexica is presented in table 4.21, an initial examination of the results identifies that positive affect is not as significant as negative affect, while *Sent* t_{-1} has a significance level of 99% across the

negative affect time series and is significant to 95% when examined against the GI positive lexicon. As has been seen previously the lexicon produced through the genetic algorithm approach has lowered the R^2 value, however the affect time series generated through the usage of particle swarm optimization is now significant to 95% compared to 90% for the other two lexica. While this is an improvement when examining the AIC weights this difference is not statistically significant to draw any definitive conclusions. Of note is the sign of the coefficient, as with the other lexica the coefficient at time $t-1$ is negative followed by a reversal. This suggests that an increase in positive affect as defined by the lexica results in a downward movement in the following day's market returns. This observation hold when examining the S&P 500 (table E.7) while with the VIX the signs are reversed with $t-1$ being positive (table E.11), this reversal of signs in the VIX analysis is apparent across all lexica.

Table 4.21: Comparison of Refined Positive Affect Against DJIA

This table presents the OLS regression results of the DJIA across negative sentiment extracted from our AP financial Newswire. Positive sentiment is produced from a raw absolute frequency count of terms as defined against the General Inquirer and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Sentiment Measurement	Positive	Positive $_{Genetic}$	Positive $_{Particle}$
Sent t_{-1}	**-23.4****	**-21.8****	***-23.3****
Sent t_{-2}	*15.6**	*13.3**	*15.0**
R^2	3.28%	3.24%	3.30%
AIC	-8727.3	-8726.7	-8727.6
$P-Value(Sent)$	*0.07**	*0.06**	**0.03****

When examined against the S&P500 there similarly seems to be little statistical significance between the inclusion of lagged positive sentiment and future market returns (significant to 90%) indeed the lexica produced through genetic algorithms results in a R^2 12% lower than that obtained from the raw lexicon, this is to the extent to which Positive *Genetics* may be excluded with a AIC weight of 0.02. As with previous evaluations there appears to be no statistical link between the affect time series and future VIX returns.

Lastly the regression analysis of the affect time series obtain from the Financial Negative *(Fin Neg)* lexicon produced by Loughran & McDonald is presented. As noted this lexicon performed somewhat poorly when evaluated in its raw form, this lexicon was produced with the domain of financial language in mind, and while developed for an examination of financial reports of a more technical nature than news reporting, not only have past studies applied this lexicon to financial news analysis, it may be assume that there would be little misclassification

as while the nature of the documents have changed the domain has not.

Table 4.22: Comparison of Refined Financial Negative Regressed Against DJIA
This table presents the OLS regression results of the DJIA across 'Financial Negative' sentiment extracted from our AP financial Newswire. 'Financial Negative' sentiment is produced from a raw absolute frequency count of terms as defined against the term list provided by Loughran & McDonald and then refined using both genetic algorithms and particle swarm. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

Sentiment Measurement	Fin Neg	Fin Neg $_{Genetic}$	Fin Neg $_{Particle}$
Sent t_{-1}	***-23.03****	***-27.49****	***-23.75****
Sent t_{-2}	*13.84**	**14.81****	12.40
R^2	3.45%	4.23%	3.66%
AIC	-8730.0	-8742.4	-8733.3
$P-Value(Sent)$	**0.025****	***0.005****	***0.01****

A significant improvement in the performance of the financial negative lexicon once refined through genetic algorithm refinement method can be seen with an R^2 increase of 22% (from 3.45% to 4.23%). While the improvement is not as apparent when the particle swarm method is applied both increase the level of statistical significance from 95% to 99%. Similarly there is some improvement examination against both S&P and VIX however not to the same extent (tables E.8 and E.12). The performance of this lexicon once refined may potentially reflect the domain specific nature (finance in this instance) and that through some refinement and improvement provides a valuable resource for the extraction sentiment within financial news once modified somewhat for the style of financial news reporting.

Discussion

This section examined the impact of feature selection on legacy lexica within the domain of sentiment analysis of financial news. Six months worth of news items and market movements were used in training the system to produce a refined word list aimed at best capturing relevant affect terms in the initial lexica. From these refined word lists the affect time series for the remaining time period could be examined. The strategy was to take a word list produced by the genetic algorithm and particle swarm optimization techniques, based on six months of news selected from a chronologically organised dataset between January 2006 and July. This analysis produced a word list which had the highest value for the fitness function and used to compute the impact of sentiment on the affect time series as a time series for the corresponding time period from January-July 2006. The impact of sentiment based on the handcrafted word lists is then compared with that produced from genetic algorithms and

particle swarm. This computation was carried out on a seven year period, July 2006-August 2012. The corresponding asset values of the time series were used in the regression. In several cases the usage of such refined lexica produce superior statistical results when compared to those obtained using the *raw* legacy lexica. A statistically significant relationship between sentiment and future market returns in the DJIA and S&P 500 has been demonstrated. For both indices an increase in sentiment (positive or negative) is followed by a downward movement of returns on the following day which is later reversed, such observations are consistent with the works of Tetlock and demonstrate that while small the impact of sentiment on market returns is statistically significant.

Through the application of machine learning a methodology has be proposed which removes terms from a given lexicon which may be misclassified within the target domain. In a number of these cases the improvement is statistically significant such that the raw lexicon may be excluded from future analysis. This improvement is not guaranteed however, in some instances the produced lexica perform worse than the raw legacy lexica. While unfortunate this is not to be unexpected, generally when the produced lexicon performs worse there appears to have been little statistical significance to begin with. When examining the frequency of positive affect terms the statistical significance is weaker (90%) when compared to negative affect frequency (95%). There is the potential that there is little underlying relationship to be extracted from lexicon refinement and the improvement obtained during training and lexicon refinement is merely an example of data mining leading to over-fitting to the training time period. A secondary issue for consideration is the size of the problem space, as noted each of the given lexica comprise a large number of features which can be problematic within biologically inspired algorithms. This could be of issue specifically when examining the stemmed version of the negative affect lexicon which has almost twice as many terms as the other given lexica. However while not consistent there is a general improvement in the results of the regression analysis indicating that the proposed methodology can deal with large feature spaces to a degree. As a means for comparison of the varying lexica against each index results are summarised in table 4.23 which outlines the co-efficient of regression for sentiment at time $t-1$, in a manner similar to Tetlock results indicating the impact of an increase in one standard deviation in sentiment against the following days market returns in basis points.

4.4.2 Refined Lexicon Content

While the resultant time-series produced from the refined dictionaries provides evidence for the success of refined lexica within the field of sentiment analysis, an examination of the resulting dictionaries themselves is also of interest. Specifically the terms which have been selected for inclusion and exclusion by the various dictionary refinement methods, an examination of

Table 4.23: A comparison of coefficients at $t-1$ for all Dictionaries and Refinement Methods
This table summarize the coefficient of $t-1$ across each of the given lexica in their raw form and refined through genetic algorithms and particle swarm optimization. These results are obtained by regressing the lagged sentiment to a max of $t-2$ across the given index, our affect time series is produced through absolute term frequency within our corpus of no exact duplicates, comprising of 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Each coefficient indicates the impact of an increase of one standard deviation of sentiment on the following market day returns in basis points (one basis point equals a return of 0.01%). Italics indicate the coefficient is significant to 90%, bold to 95% while bold and italics indicates the coefficient is significant to 99%. Instances where the impact is greater than that occurring from an increase of one standard deviation of market returns are underlined.

		DJIA	S&P 500	VIX
Negative	Negative	*-26.82***	*-27.91***	**73.09****
	Genetics	*-28.96***	*-30.78***	**60.41****
	Particle	*-29.51***	*-30.53***	**71.02****
Negative-Inf	Negative-Inf	*-26.66***	*-27.94***	73.32
	Genetics	*-28.50***	*-26.43***	65.60
	Particle	*-28.75***	*-29.93***	79.06
Positive	Positive	**-23.46****	**-24.08****	54.83
	Genetics	**-21.77****	*-16.69**	46.77
	Particle	*-23.31***	**-22.85****	57.57
Fin Neg	Fin Neg	*-23.03***	*-24.27***	**67.75****
	Genetics	*-27.49***	*-25.57***	**67.84****
	Particle	*-23.75***	*-28.63***	**77.55****

such terms may provide insight into the nature of language in financial reporting, previous works have stated that when examining sentiment analysis lexicon development based on a corpus examination may be superior that that developed purely from intuition and prior knowledge [81].

Henry & Leone [37] provide an extensive examination between a domain specific lexicon and that given through a general language lexicon; in this instance the General Inquirer. They examine not only the improvements in results when examining affect term frequency within the corpus used by Tetlock [104] but also an examination in terms included and excluded across the varying dictionaries. A specific example is that of the term *'division'* while in general language such a term is indeed negative in this instance the condition of disagreement or antagonism between two groups, within financial news such a term would most likely simply refer to varying departments or branches of an organisation. So as to evaluate the performance of the system and investigate the produced lexica table 4.24 provides a sample of terms selected for exclusion and inclusion from the GI Dictionary *Negative* category, this dictionary is obtained through training the system against the Dow Jones Industrial Average over a six month period

in 2006 from the given dataset.

Table 4.24: Sample of Include and Exclude Terms obtained from the Pruned Negative GI Lexicon.

Exclude	Include
abolish, angry, argue	badly, bane, contagious
bar, bail, bribe	chaos, contamination, crisis
capital, chase, clash	demise, distress, fallout
competition, complex, concern	handicap, hectic, liquidation
division, drop, fail	misinform, pain, peril
tax, pound, push	pretense, rebut, rat

At this instance it may be of interest to examine and discuss some of the terms and the reasoning for their inclusion/exclusion. The exclusion of the term 'division' is promising as this was a key term identified by Henry & Leone stating how while a division is indeed negative when it is applied to a disagreement between two individuals, with finance this term will simply reference different departments within a single company. Capital within negative affect is most likely referring towards capital punishment, yet within financial or economic terminology is used in reference to one of the factors of production, money or assets used within the course of production. Pound similarly is a violent negative term meaning beating yet here it can expect its usage is for discussion of the British currency. Lastly Tax while a negative term to most individuals and indeed is relevant within financial news may provide no information without future contextual information, taxes may be discussed in terms of government levies, company *Pre-Tax Profits* or numerous other instances, this demonstrates that while negative and relevant within financial language is not in itself a negative affect term within the domain. However there are also terms which would not intuitively be expected for excluding a relevant in the field of financial news. Terms such as *bride* and *bail* are both negative and it could be assumed that there is no possibility these could be used as anything other than negative in financial and economic reporting. This indicates that while the proposed methodology does excluded a number of terms which may be misclassified when examining financial news, there is no guarantee that all will be excluded. When applying Genetic Algorithms to problem areas with a large number of features problems can emerge, in this instance the evaluating the Negative GI dictionary results in 2,004 individual features with one for each term in the word list.

Regarding terms selected for inclusion within the refined lexicon; contagious and contamination are both note worthy terms, while the origin of these was towards the spread of disease and impurity due to a foreign substance both are used in a metaphorical sense within financial and economic news most recently when discussing the risk of debt influencing the economies

Table 4.25: Sample of Include and Exclude Terms obtained from the Pruned Positive GI Lexicon.

Exclude	Include
aid, asset, aware	acquit, baptism, buy
boom, boost, capitalize	clout, efficacy, embrace
commission, company, council	exuberant, flirt, good
credit, interest, partnership	infer, mobilize, modest
pay, share, train	outstanding, peace ,rally

of other countries.

> *U.S. Banks Face Contagion Risk From Europe Debt- Bloomberg November* 17^{th} *2011*

These provide an example of the expressive and metaphorical nature of language; financial news style is not without its adoption of various domains' terminology for metaphorical usage. While terms such as chaos and crisis, are both highly emotional terms with significant negative bearing both frequently used for the description of large decreases with economies and markets.

While the system has excluded terms that appear intuitive there are term selected for inclusion that are questionable such as *handicap& rat*, an examination of the training corpus identifies that such terms never occur as would be expected; thus their inclusion has no bearing on the fitness function however demonstrate the potential value of implementing a penalty function based on solution complexity as the if there is a later usage of such terms may prove detrimental to analysis.

Henry & Leone provide a list of the 20 most frequent terms which are not categorised as negative within a specialised dictionary yet are within the GI;

expense, taxes, cost, costs, tax, press, loss, differ, capital, depreciation, services, excluding, service, services, charges, charge, short, limited , foreign, competitive, involve

The authors argue that such terms are misclassified when examining terminology when applied to the financial domain which demonstrates issues with the usage of the raw negative affect lexicon.

Of this word list the only terms not selected for exclusion are *'service'* and *'services'*. These results are encouraging as they provide a benchmark by which to compare the system; wherein expert have manually examined a comprehensive lexicon and made refinements based on their own understanding of the terminology which is similar to those produced through the system, however in an automated manner.

Table 4.25 provides a summary of terms selected for inclusion and exclusion from the positive affect category of the GI when refined using the GA refinement on the DJIA series. When examining the output again a number of terms selected for inclusion such as *'flirt'*, *'modest'* and *'clout'* would appear to have no meaning within financial news are again have no occurrences within the training data and as such their inclusion has so bearing on the system's performance. By contrast the exclusion of terms such as *'share'*, *'company'*, *'credit'* and *'interest'* is beneficial, these four terms are extremely common within financial literature, the term share when applied to stock markets is in itself a completely neutral term with no affect bearing. Similarly to company, credit and interest; the system appears to have successfully excluded such terms which here are merely financial concepts.

'Rally' when used with this domain would be applied to *"A period of sustained increases in the prices of stocks, bonds or indexes. "*[2] again a term which when manually examined does have a positive affect when applied to the financial domain. The inclusion of the term *exuberant* is striking, within economic commentary this term is most famously associated with a speech by Alan Greenspan then chairman of the Federal Reserve Board in Washington's usage of in the question

> *"But how do we know when irrational exuberance has unduly escalated asset values, which then become subject to unexpected and prolonged contractions as they have in Japan over the past decade?"-December 1996*

This phrase has been defined as "Unsustainable investor enthusiasm that drives asset prices up to levels that aren't supported by fundamentals" [3]. This concept and the potential impact of this single statement was later explored regarding the dot-com bubble and its subsequent burst by Schiller [89].

Henry & Leone provide a similar list of common positive terms within the GI which are not found within their domain specific lexicon:

share, company, forward, shares, interest, equity, basic, actual, common, contact, outstanding, consolidated, commission, value, call, primarily, paid, gain, ability.

As with the negative affect category the system has eliminated all but two of these terms *"common"* and *"outstanding"* again suggesting the successful ability of the system to accurate automated lexicon refinement reaching similar results to those compiled through manual examination.

Lastly a summary of the refined Financial Negative lexicon provided by Loughran & McDonald is presented; as noted this lexicon differs from the others in that it was produced through expert knowledge and while aimed towards financial reports rather than news it does however provide a valuable resource. A summary examination of table 4.26 provides some

[2] http://www.investopedia.com/terms/r/rally.asp

[3] http://www.investopedia.com/terms/i/irrationalexuberance.asp

Table 4.26: Sample of Include and Exclude Terms obtained from the Refined Financial Negative Lexicon.

Exclude	Include
abuse, adverse, bailout	aberrant, abused, accuses
bankrupted, closed, costly	allege, annuls, anomalies
defraud, devalue, embarrass	bankruptcy, bribery, caution
fraud, late, loss	default, defer, dismal
omit, resign, shut	fail, foreclosures, liquidations

interesting analysis, a number of terms within the Financial Negative Lexicon are in reference to legal concepts such as *'Fraud'* and inter company issues such as *'Resign'*. Within an individual company or organisation the resignation or act of fraud of a senior employee could be assumed to indeed have a major impact on the companies returns. However as we are examining sentiment on a much wider level such instances would not in themselves impact greatly on the economy or overall market index as a whole. The exclusion of a term such as closed is of note as when applied to the economy as a whole this term may be frequently used in reference to the closing value of a market index, yet in a company's financial reports this could be in reference to the closure of offices. Whereas we can see that terms such as bankruptcy and default may have a wider impact on the economy. The exclusion of the term bailout is noteworthy; our natural opinion would be such that this term would be of key issue and its exclusion is unfortunate however this may be due to the time-frame across which the system was trained. The importance of bailouts has become more common during the period of the financial crisis and is potentially absent from the time period on which the system was trained. Alternatively this may simply be another example where the complexity of the system and wide number of variables demonstrates the failure of genetic algorithms and particle swarm optimization regarding a thorough problem space exploration.

We summaries the refined lexica sizes in table 4.27

Table 4.27: Refined Lexica Size.

Lexicon	Raw Lexicon	Refined Lexicon	% Eliminated
Negative	2004	785	60.8%
Negative-Inf	4187	1762	57.9%
Positive	1634	617	62.2%
Negative Financial	2337	906	61.2%

As is visible each instance results in a lexica refined by approximately 60% while improving

Table 4.28: The Top 30 Terms by $Tf - Idf$ score from a sample of the corpus

Term		
yen	percentage	sales
dollar	oil	trade
index	bank	billion
cents	fell	million
rates	rate	rose
points	euro	stock
shares	prices	currency
quarter	spending	European
he	countries	Japanese
interest	US	treasury

regression analysis in a number of instances. This demonstrates as has been proposed that through some process it is possible to eliminate a substantial number of terms from a legacy lexicon towards the production of a domain specific contemporary lexicon while improving textual analysis results. The above results are those obtained from applying genetic algorithms for lexica refinement by regressing the affect time series against the DJIA returns.

4.5 Keyword Identification

Following is a short evaluation of key term identification within this corpus using a term frequency inverse document frequency ($tf - idf$)approach. As mentioned in section 2.3.2 by evaluating the frequency of a term in a document and examining it against how many documents it is found in across the entire corpus a list of key terms can be generated. An advantage of $tf - idf$ is that while easy to calculate it has been shown to perform rather well when it comes to identifying key terms in a document [85]. As this corpus was collected for sentiment analysis of financial news it would be expected that a $tf - idf$ measurement would return a large number of terms related to the financial world. Terms with a high $tf - idf$ score are said to be important in describing documents, while those with low scores occur many times across all documents, to an extent that they possibly provide little descriptive information, these term will frequently be so called *Stop Words* as mentioned in section 2.3.2.

A number of the terms in table 4.28 are ones that would be expected to be important in a news corpus on financial news concerning the US economy. Several of these terms most likely relate to stock markets, terms such as *trade, index, fell, rose, stock* and *countries*. Through the

analysis of the frequency of terms within a corpus compared to the number of documents it term is in provides a method for the automated generation of key terms in the given domain. If the terms with the lowest $Tf - Idf$ scores are examined many are terms such as *to, a, the, of*, as previously mentioned these are common stop terms who's frequency is found across all documents the idf score is sufficiently low that they can be supposed to provide little information about the key topics in the corpus.

4.6 Findings and Chapter Discussion

Within this chapter we have examined the impact of sentiment within financial news on future market movements, the impact of duplicates and the effectiveness of automated refinement of legacy lexica as the main areas of focus.

Following an examination on the configuration in regards to weighting and the impact of duplicates we examined the impact of varying reference lexica when regressed against the DJIA, S&P 500 and VIX. The initial results showed that similarly to past studies negative affect was shown to have the greatest level of statistical significance when compared to the positive affect time series. While both negative affect series obtained from the negative lexicon were statistically significant to 95% within the DJIA and S&P 500 this was no observed within the positive affect time series. Similarly there appears to be little statistical evidence of a link between corpus sentiment and movement in the VIX. Interestingly the expert lexicon examined performed poorly; while noted that this lexicon was compiled for an examination of negative affect within corporate reports it would be expected that there would be less misclassification within this lexicon.

Having examined the impact of varying lexicon against the three proposed indices we observe statistical significance only at the point $t-1$. The regression model is then simplified to include a maximum lag of $t-2$, which leads to an increase in the level of statistical significance from 95% to 99%. We observe that while the inclusion of a lag at $t-2$ is significant to 95% in the negative affect models regressed against future market returns.

We observe that in a number of regression models the usage of the trained refined lexicon outperform the results obtained from the initial legacy lexicon. While there is not a consistent improvement obtained from the refined lexica as noted due to the number of features obtained from the reference lexica. Secondly in some of the instances there appears to have been little statistical significance when examining the any of the affect time series against VIX returns.

The secondary method for an evaluation of the effectiveness of the system to refine lexica is through a manual examination of the produced lexica content. We observe a number of terms which have been identified as being misclassified within financial news yet have obtained these through an objective machine learning technique rather than through subjective manual

examination. We observe that the developed system successfully removes approximately 60% of terms from the raw lexica while generally improving performance accuracy. These reductions may be increase given that as noted a number of terms were not selected for exclusion due to them never having occurred within the corpus.

Chapter 5

Closing Remarks and Future Work

The research presented in this thesis has examined the implementation of biological algorithms which allow for the automated refinement of a given lexicon within the field of sentiment analysis as applied to financial news. The following chapter presents a discussion of some of the key learning outcomes and achievements and proposes open areas of future work.

5.1 Contributions

The primary goals within this research were an examination of sentiment within financial news and producing a system allowing for the refinement of a given lexicon to better extract sentiment within the given domain. The developed system allows for the large scale collection a refinement of texts for corpus development. With an collection of news items the corpus can be examined for collisions and a time series can be generated using a given word list3.2.2. This systematic nature allows for the generation of affect time series in a manner requiring minimal human interaction.

A key topic within this work has been firstly the importance of the domain towards which the reference lexicon of affect terms is being applied. Secondly that while valuable; the subjective nature of lexicon developed from interviews or researcher's knowledge of the domain present problems. The automated manner by which the system evaluates a reference lexicon and refines or eliminates terms by evaluating against the concurrent financial market provides a systematic approach to this problem. The financial movements provide an external source of information from which we may evaluate the system.

The results obtained when examining the performance of the developed system against alternative lexica are encouraging. While by no means perfect the produced refined lexica frequently outperformed the initially presented raw lexica, and display a high level of statistical significance when examined against financial indices. This supports past findings that when examining a legacy lexicon against a given domain the correct affect identification of terms is

not guaranteed. The system appears to successfully allow for the refinement of a given lexicon through the elimination of term which within the field of financial news have a different affect bearing.

Furthermore when examining the terms selected for inclusion and exclusion we witnessed a number of instances where the same terms had been identified as those chosen through manual evaluation. This is a secondary benefit of lexicon refinement in that by examining the resulting refined lexicon we may develop a better understanding of the nature of financial reporting language. Through an examination of terms selected for exclusion and inclusion presented in section 4.4.2 we may gain insight into the nature of financial language. While the exclusion of *pound & capital* from the negative affect dictionary would appear intuitive, excluding *fail & drop* would not necessarily be identified from manual examination. We may suppose that these terms do not provide adequate information on their own but rather the surrounding context must be taken into consideration.

An evaluation of the impact of duplicates was a secondary outcome. While initially developed to sanitize the corpus to account for duplicates news items which in themselves may contain little to no new information; experimental analysis indicates that the inclusion of such items has a significant impact on results when examining sentiment against market returns. We may suppose that the repetition of a news even indicates its importance, similarly to the repetition of a term with a domain indicates its acceptance and importance within domain's language. When news items are released which speculate on upcoming announcements are followed with a near identical piece with modifications being a summary of the announcement it should appear that the early item contains no new information. However this may indicate the level of importance of the topic being discussed.

5.2 Future Work

During the course of research a number of areas of future potential research arise. The four main areas of future work are outlined following.

- Firstly as noted the examination of collision provided some interesting results. The impact of such duplicates was an unexpected outcome; future work may examine the frequency of which these news items occur and their nature. During a period of financial turmoil we may expect a large number of collisions are news items are being continuously re-released with minor modifications containing updated market values. This would be consistent with modern reporting; a continuously updated stream of information.

- While an examination of the frequency of such collisions is of interest the nature of such collisions is also an area of potential future research. In section 2.4.2 we presented a

sample of identified collisions with a sample corpus. Within one there was a tonal shift with the movement from a rise in oil prices due to growth towards a rise resulting from potential conflict with Iran. An exploration of collisions would allow for a examination of how a news item changes over time. Do authors frequently alter the tonal shift of a news item for the purpose of sensationalism while providing no new information?

- An additional presented collision indicated a shift from a speculative news item *Stocks Set to Open Down on Oil Prices* to a conformation later in the day *Stocks Open Lower on Oil Prices* along with a confirmation regarding the nature of an upcoming report due for release. A relaxing of the criteria for a collisions would allow for the development of a system allowing for the tracking of discussion of an event over time. Rather than identifying duplicate news items it may be possible to identify news items discussing the same topic.

- The origin and nature of the provided lexicon is irrelevant, regardless of the nature of the terms the system should be able to refine any given list removing irrelevant or misclassified terms. Therefore it may be of interest to integrate this system with some form of corpus based approach to lexicon development. A reference lexicon composed of a number of terms with the highest frequency of occurrences within a training corpus could be provided to the system and potential valuable insight may be obtained. Alternatively a weirdness based approach could be implemented, the returned terms can be concluded to being important within a given domain based on the higher level of occurrences when compared to general language. If these terms are important to the given domain the refined lexicon should contain affect term which are important within the given domain.

Lastly as discussed while this work was developed towards an examination of sentiment in financial news the domain should in theory be irrelevant. Provided there is some available performance measurement function the system should be able to produce a refined lexicon.

Appendix A

Impacting of Weighting Analysis

Table A.1: Impact of Weights When Regressing Against S&P 500.

This table presents the OLS regression results of the S&P 500 across negative sentiment extracted from our AP financial Newswire corpus containing no exact Duplicates. We compare the impact of three weighting functions Absolute negative term frequency, relative term frequency and $tf - idf$ as defined by the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations we use Newey and West standard errors so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

	Absolute Frequency	$Tf - Idf$	Relative Frequency
Sent t_{-1}	***-30.0****	***-29.5****	**-8.9****
Sent t_{-2}	12.9	12.8	5.9
Sent t_{-3}	4.1	3.6	-3
Sent t_{-4}	11	10.4	7.9
Sent t_{-5}	-6.9	-7.5	-6.8
S&P 500$t-1$	***-21.5****	***-21.6****	***-21.1****
S&P 500$t-2$	-13.3	-13.2	-12.8
S&P 500$t-3$	3.7	3.6	3.2
S&P 500$t-4$	2.1	-2.2	-1.8
S&P 500$t-5$	8.7	-9.1	-8.2
R^2	4.27%	4.31%	3.07%
AIC	-8427.8	-8428.4	-8402.3
P-Value(Sent)	**0.05****	**0.04****	0.15

Table A.2: S&P AIC Weights Analysis

This table presents an evaluation of the different weighting functions applied to negative affect against the S&P 500. Negative affect is extracted from the AP Financial Newswire corpus with no exact duplicates. Our regression model includes lags of $t-1$ to $t-5$ of both sentiment and market returns. Results in italics indicate the model may be excluded from further analysis.

Weighting Method	*AIC*	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
Tf-Idf	-8428.43	0.00	1.00	0.58
Abs	-8427.79	0.64	0.73	0.42
Relative	*-8402.26*	*26.16*	*0.00*	*0.00*

Table A.3: Impact of Weights When Regressing Against VIX.

This table presents the OLS regression results of the VIX across negative sentiment extracted from our AP financial Newswire corpus containing no exact Duplicates. We compare the impact of three weighting functions Absolute negative term frequency, relative term frequency and $tf-idf$ as defined by the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations we use Newey and West standard errors so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

	Absolute Frequency	$Tf-Idf$	Relative Frequency
Sent t_{-1}	**48.7****	**84.6****	19.4
Sent t_{-2}	-21.3	**-37.9****	**-55.9****
Sent t_{-3}	1.3	**7.5****	13.8
Sent t_{-4}	-25.8	**-45.9****	-7.0
Sent t_{-5}	1.9	**5.6****	15.5
VIX $t-1$	***-104.6*****	***-104.8*****	***-104.9*****
VIX $t-2$	***-66.7*****	**-66.9****	**-59.2****
VIX $t-3$	**-53.2****	**-53.2****	**-48.0 ****
VIX $t-4$	*-36.1**	*-36.5**	*-33.9**
VIX $t-5$	**-50.0*****	**-50.6****	**-49.3****
R^2	3.68%	3.70%	3.46%
AIC	-3719.9	3720.3	-3716.4
P-Value(Sent)	*0.13*	0.13	0.19

Table A.4: VIX AIC Weights Analysis

This table presents an evaluation of the different weighting functions applied to negative affect against the VIX. Negative affect is extracted from the AP Financial Newswire corpus with no exact duplicates. Our regression model includes lags of $t-1$ to $t-5$ of both sentiment and market returns. Results in italics indicate the model may be excluded from further analysis.

Weighting Method	*AIC*	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
Tf-Idf	-3720.32	0.00	1.00	0.50
Abs	-3719.96	0.66	0.83	0.42
Rel	*-3716.48*	*3.84*	*0.15*	*0.07*

Appendix B

Impact of Duplicates Analysis

Table B.1: Impact of Duplicates When Regressing Against S&P 500.

This table presents the OLS regression results of the S&P 500 across negative sentiment extracted from our AP financial Newswire corpora examining the impact of duplicates. Negative affect is constructed based on absolute term frequency of negative terms as defined against the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 we use Newey and West standard errors so as to account for autocorrelation, italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

	No Duplicates	No Exact Duplicates	Raw Corpus
Sent t_{-1}	-13.1	***-30******	***-29.2*****
Sent t_{-2}	9.9	12.9	13.2
Sent t_{-3}	-4.3	4.1	5
Sent t_{-4}	11.8	11.1	*9.6*
Sent t_{-5}	-14.4	-6.9	-6.3
S&P 500 $t-1$	***-21.1*****	***-21.5*****	***-21.2*****
S&P 500 $t-2$	-13.2	-13.3	-13.1
S&P 500 $t-3$	2.8	3.7	4.1
S&P 500 $t-4$	-2.4	2.1	-1.9
S&P 500 $t-5$	-8.6	8.7	-8.5
R^2	3.41%	4.27%	4.25%
AIC	-8414.17	-8427.79	-8427.43
$P-Value(Sent)$	0.24	**0.05****	**0.04****

Table B.2: S&P 500 Duplicates AIC Analysis

We calculate the AIC weights to evaluate the effect of the exclusion of duplicates and near duplicates within our corpus. These results are obtained from an examination of absolute negative affect frequency within the three AP financial Newswire corpora, with negative terms identified according to the GI regressed against S&P 500 returns. Results in italics indicate that the relative weight of the model indicates that the model can be excluded.

	AIC	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
No Exact Duplicates	-8427.79	0.00	1	0.54
Raw Corpus	-8427.43	0.36	0.83	0.45
No Duplicates	*-8414.17*	*13.62*	*0.00*	*0.00*

Table B.3: Impact of Duplicates When Regressing Against VIX.

This table presents the OLS regression results of the VIX across negative sentiment extracted from our AP financial Newswire corpora examining the impact of duplicates. Negative affect is constructed based on absolute term frequency of negative terms as defined against the General Inquirer. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 we use Newey and West standard errors so as to account for autocorrelation, italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***).

Basis Points	No Duplicates	No Exact Duplicates	Raw Corpus
Sent t_{-1}	27.6	**85.3****	**81.2****
Sent t_{-2}	-5.4	-37.3	-38.9
Sent t_{-3}	8.6	2.2	-1.8
Sent t_{-4}	-39.9	-45.2	-35.4
Sent t_{-5}	15.6	3.3	1.8
VIX $t-1$	***-104.8*****	***-104.6*****	***-104.2*****
VIX $t-2$	**-64.0****	**-66.7****	**-66.0****
VIX $t-3$	**-54.9****	**-53.2****	**-52.4****
VIX $t-4$	*-36.0**	*-36.1**	*-35.6**
VIX $t-5$	*-49.5**	**-50.0****	**-49.6****
R^2	3.19%	3.68%	3.19%
AIC	-3712.33	-3719.96	-3712.33
$P-Value(Sent)$	0.84	*0.01**	0.84

Table B.4: VIX Duplicate Impact AIC Calculations

We calculate the AIC weights to evaluate the effect of the exclusion of duplicates and near duplicates within our corpus. These results are obtained from an examination of absolute negative affect frequency within the three AP financial Newswire corpora, with negative terms identified according to the GI regressed against VIX returns. Results in italics indicate that the relative weight of the model indicates that the model can be excluded.

	AIC	Δ_i	$exp(-0.5 * \Delta_i)$	W_i
No Exact Duplicates	-3719.96	0.00	1.00	0.57
Raw Corpus	-3719.33	0.63	0.73	0.42
No Duplicates	*-3712.33*	*7.63*	*0.02*	*0.01*

Appendix C

Varying Lexica Performance

The following tables present the results of OLS regression results cross all dictionaries against returns for each of the three indices. Sentiment is calculated as absolute term frequency obtained from our corpus of no exact duplicates. Regression models include a lagged sentiment values from $t-1$ to $t-5$, similarly lagged values of market returns are also included in the regression model. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Table C.1: Regression of Varying Dictionaries against DJIA.

DJIA-Basis Points	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	***-29.2***	***-29.2***	***-28.9***	***-25.8***
Sent t_{-2}	13.5	14.8	8.4	11.1
Sent t_{-3}	6.0	5.8	11.8	5.0
Sent t_{-4}	8.5	6.1	5.5	2.6
Sent t_{-5}	-6.2	-4.7	-3.1	-0.9
R^2	4.57%	4.53%	4.24%	4.14%
AIC	-8716.52	-8715.89	-8711.31	-8709.78
$P-Value(Sent)$	**0.05****	*0.06**	0.138	0.13

Table C.2: Regression of Varying Dictionaries against S & P 500.

S&P-Basis Points	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	***-30.01******	***-30.36******	**-29.36****	**-27.46*****
Sent t_{-2}	12.91	14.30	7.02	9.83
Sent t_{-3}	4.13	4.01	11.01	3.52
Sent t_{-4}	11.05	8.32	8.09	4.56
Sent t_{-5}	-6.87	-5.12	-4.42	-0.21
R^2	4.27%	4.24%	4.0%	3.9%
AIC	-8427.79	-8427.253	-8422.82	-8422.41
P-Value(Sent)	**0.05****	*0.061**	0.15	0.14

Table C.3: Regression of Varying Dictionaries against VIX.

VIX-Basis Points	Negative	Negative-Inf	Positive	Fin Neg
Sent t_{-1}	**85.3****	**87.6****	*75.6**	**88.2*****
Sent t_{-2}	-37.5	-37.5	-16	-26.5
Sent t_{-3}	22.25	6.1	-3.9	-5.7
Sent t_{-4}	-45.2	-45.9	-43.6	-31.8
Sent t_{-5}	3.3	-1.7	-1.7	-17.1
R-Squared	3.68%	3.71%	3.50%	3.70%
AIC	-3719.96	-3720.53	-3717.23	-3720.27
P-Value(Sent)	0.11	*0.09**	0.4	*0.084**

Appendix D

Refined Lexica Analysis

The following tables present the results of OLS regression results cross all dictionaries refined using two feature selection methods genetic algorithms and particle swarm optimization against returns for each of the three indices applying the two feature selection methods genetic algorithms and particle swarm optimization. Sentiment is calculated as absolute term frequency obtained from the corpus of no exact duplicates. Regression models include a lagged sentiment values from t - 1 to t - 5, similarly lagged values of market returns are also included in the regression model. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, we use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Table D.1: Negative DJIA Refined Lexica

	Negative	Negative $_{Genetics}$	Negative $_{Particle}$
Sent t_{-1}	***-29.2***	***-34.6***	***-32.3***
Sent t_{-2}	13.5	12.6	*16.8**
Sent t_{-3}	6.0	6.1	4.4
Sent t_{-4}	8.5	9.9	6.9
Sent t_{-5}	-6.2	-2.0	-3.5
Dow t_{-1}	***-19.3***	***-19.9***	***-19.2***
Dow t_{-2}	-11.3	-11.5	-11.2
Dow t_{-3}	5.3	5.4	5.1
Dow t_{-4}	-2.2	-1.8	-2.2
Dow t_{-5}	-9.4	-9.8	-9.3
R^2	4.57%	5.23%	4.79%
AIC	-8716.52	-8727.171	-8720.108
P-Value(Sent)	**0.05****	**0.03****	**0.02****

Table D.2: Negative-Inf DJIA Refined Lexica

	Negative-Inf	Negative-Inf $_{Genetics}$	Negative-Inf $_{Particle}$
Sent t_{-1}	***-29.2******	***-31.1******	***-31.6******
Sent t_{-2}	14.8	**17.4****	*16.7**
Sent t_{-3}	5.8	3.2	7.8
Sent t_{-4}	6.1	5.5	3.7
Sent t_{-5}	-4.7	-2.7	-4.0
Dow t_{-1}	***-19.5******	***-19.5******	***-19.9******
Dow t_{-2}	-11.5	-11.3	-11.5
Dow t_{-3}	5.3	5.2	5.2
Dow t_{-4}	-2.0	-2.1	-2.0
Dow t_{-5}	-9.5	-9.7	-9.6
R^2	4.53%	4.94%	4.77%
AIC	-8715.88	-8722.48	-8719.81
P-Value(Sent)	**0.05****	**0.02****	**0.05****

Table D.3: Positive DJIA Refined Lexica

	Positive	Positive$_{Genetics}$	Positive $_{Particle}$
Sent t_{-1}	**-28.8****	***-27.8******	***-29.4******
Sent t_{-2}	8.4	3.4	5
Sent t_{-3}	11.8	14.0	13.1
Sent t_{-4}	5.5	*15.3**	11.3
Sent t_{-5}	-3.1	*-11.1**	-5.8
Dow t_{-1}	***-18.5******	***-18.4******	***-18.6******
Dow t_{-2}	-9.9	-8.9	-9.6
Dow t_{-3}	4.5	5.4	4.8
Dow t_{-4}	-2.7	-3.5	-2.8
Dow t_{-5}	-10.2	-10.4	-10.1
R^2	4.24%	4.77%	4.52%
AIC	-8711.314	-8719.768	-8715.859
$P-Value(Sent)$	0.13	**0.01****	**0.036****

Table D.4: Financial Negative DJIA Refined Lexica

	Fin Neg	Fin Neg$_{Genetics}$	Fin Neg$_{Particle}$
Sent t_{-1}	***-25.84******	***-31.18******	***-27.40******
Sent t_{-2}	11.12	10.29	7.98
Sent t_{-3}	4.99	7.88	6.81
Sent t_{-4}	2.55	0.71	4.62
Sent t_{-5}	-0.92	1.36	-1.35
Dow t_{-1}	***-19.74******	***-19.77******	***-19.84******
Dow t_{-2}	-11.81	-11.34	-11.78
Dow t_{-3}	4.59	3.78	4.57
Dow t_{-4}	-1.94	-2.53	-1.93
Dow t_{-5}	-9.58	-10.23	-9.69
R^2	4.14%	5.06%	4.48%
AIC	-8709.777	8724.395	-8715.19
$P-Value(Sent)$	0.13	**0.0432****	*0.1*

Table D.5: Negative S&P 500 Refined Lexica

S&P	Negative	Negative Genetics	Negative PSO
Sent t_{-1}	***-29.98******	***-36.98******	***-36.25******
Sent t_{-2}	12.90	*15.33**	14.89
Sent t_{-3}	4.13	5.60	2.24
Sent t_{-4}	11.04	6.22	12.40
Sent t_{-5}	-6.87	1.03	-2.05
S&P 500 t_{-1}	***-21.53******	***-21.66******	***-21.69******
S&P 500 t_{-2}	-13.29	*-13.14**	*-13.81**
S&P 500 t_{-3}	3.72	3.82	3.68
S&P 500 t_{-4}	-2.08	-1.63	-1.72
S&P 500 t_{-5}	-8.65	-9.42	-8.31
R^2	4.27%	4.84%	4.59%
AIC	-8427.79	-8436.88	-8432.94
$P-Value(Sent)$	**0.05****	***0.01******	**0.02****

Table D.6: Negative-Inf S&P 500 Refined Lexica

	Negative-Inf	Negative-Inf Genetics	Negative-Inf PSO
Sent t_{-1}	***-29.2******	***-31.1******	***-31.6******
Sent t_{-2}	14.8	**17.4****	*16.7**
Sent t_{-3}	5.8	3.2	7.8
Sent t_{-4}	6.1	5.5	3.7
Sent t_{-5}	-4.7	-2.7	-4.0
S&P 500 t_{-1}	***-19.5******	***-19.5******	***-19.9******
S&P 500 t_{-2}	-11.5	-11.3	-11.5
S&P 500 t_{-3}	5.3	5.2	5.2
S&P 500 t_{-4}	-2.0	-2.1	-2.0
S&P 500 t_{-5}	-9.5	-9.7	-9.6
R^2	4.53%	4.94%	4.77%
AIC	-8715.89	-8722.48	-8719.82
$P-Value(Sent)$	*0.057**	**0.02****	*0.064**

Table D.7: Positive S&P 500 Refined Lexica

	Positive	Positive Genetics	Positive PSO
Sent t_{-1}	**-29.37****	*-15.51**	**-31.11****
Sent t_{-2}	7.02	0.25	1.60
Sent t_{-3}	11.01	3.47	15.90
Sent t_{-4}	8.10	**14.66**	15.15
Sent t_{-5}	-4.43	*-8.12*	-10.47
S&P 500 t_{-1}	***-20.77******	***-20.36******	***-21.15******
S&P 500 t_{-2}	-11.74	-11.49	-11.47
S&P 500 t_{-3}	3.07	3.39	3.54
S&P 500 t_{-4}	-2.60	-2.72	-2.77
S&P 500 t_{-5}	-9.64	-9.52	-9.78
R^2	3.96%	3.87%	4.36%
AIC	-8422.82	-8421.34	-8429.22
$P-Value(Sent)$	0.149	**0.025****	*0.06**

Table D.8: Financial Negative S&P 500 Refined Lexica

	Fin Neg	Fin Neg Genetics	Fin Neg PSO
Sent t_{-1}	***-27.46****	***-24.99****	***-29.64****
Sent t_{-2}	9.83	7.40	12.37
Sent t_{-3}	3.52	1.65	6.81
Sent t_{-4}	4.56	5.15	3.43
Sent t_{-5}	-0.21	2.50	-1.82
S&P 500 t_{-1}	***-22.13****	***-21.67****	***-22.09****
S&P 500 t_{-2}	*-13.87**	*-13.62**	*-13.40**
S&P 500 t_{-3}	2.86	2.77	3.42
S&P 500 t_{-4}	-1.81	-1.71	-1.61
S&P 500 t_{-5}	-8.89	-8.54	-9.01
R^2	3.93%	4.23%	4.40%
AIC	-8422.41	-8427.08	-8429.91
$P-Value(Sent)$	0.14	*0.10**	**0.05****

Table D.9: Negative VIX Refined Lexica

	Negative	Negative $_{Genetics}$	Negative $_{Particle}$
Sent t_{-1}	**85.23****	*74.50**	**98.56****
Sent t_{-2}	-37.29	-53.65	-37.32
Sent t_{-3}	2.25	57.16	17.95
Sent t_{-4}	-45.12	**-77.46****	*-62.80**
Sent t_{-5}	3.32	16.85	-6.69
VIX t_{-1}	***-104.64****	***-101.09****	***-105.01****
VIX t_{-2}	**-66.69****	**-66.33****	**-66.24****
VIX t_{-3}	**-53.21****	**-51.20****	**-53.60****
VIX t_{-4}	*-36.06***	*-37.02***	*-35.95***
VIX t_{-5}	**-50.00****	**-49.49****	**-50.33****
R^2	3.68%	3.73%	3.73%
AIC	-3719.96	-3720.86	-3720.75
$P-Value(Sent)$	0.111	0.124	*0.087**

Table D.10: Negative-Inf VIX Refined Lexica

	Negative-Inf	Negative-Inf $_{Genetics}$	Negative-Inf $_{Particle}$
Sent t_{-1}	**87.69****	**75.21****	**89.90****
Sent t_{-2}	-37.53	-36.23	-40.19
Sent t_{-3}	6.06	5.78	9.86
Sent t_{-4}	-45.99	-46.99	*-51.86*
Sent t_{-5}	-1.73	14.20	1.55
VIX t_{-1}	***-105.23****	***-103.50****	***-104.04****
VIX t_{-2}	***-67.60****	**-64.80****	**-67.12****
VIX t_{-3}	**-53.34****	**-52.79****	**-52.63****
VIX t_{-4}	*-35.86**	*-35.49**	*-35.40**
VIX t_{-5}	**-50.14****	**-51.20****	**-50.22****
R^2	3.71%	3.64%	3.82%
AIC	-3720.53	-3719.45	-3722.30
$P-Value(Sent)$	*0.091**	0.193	*0.07**

Table D.11: Positive VIX Refined Lexica

	Positive	Positive $_{Genetics}$	Positive $_{Particle}$
Sent t_{-1}	*103.43**	90.26	**125.61****
Sent t_{-2}	-21.84	-21.26	-17.42
Sent t_{-3}	-5.36	-9.93	-8.44
Sent t_{-4}	-59.64	**-98.44****	*-81.49**
Sent t_{-5}	-2.45	54.27	-5.24
VIX t_{-1}	***-102.96****	***-101.97****	***-103.18****
VIX t_{-1}	**-63.21****	**-61.75****	**-63.34****
VIX t_{-1}	**-54.68****	**-52.97****	**-54.68****
VIX t_{-1}	*-37.29**	*-37.91**	*-37.38**
VIX t_{-1}	**-52.39****	**-50.98****	**-52.92****
R^2	3.50%	3.57%	3.64%
AIC	-3717.23	-3718.32	-3719.38
$P-Value(Sent)$	0.404	0.252	0.21

Table D.12: Financial Negative VIX Refined Lexica

	Fin Neg	Fin Neg *Genetics*	Fin Neg *Particle*
Sent t_{-1}	***162.51****	***138.83****	***149.75****
Sent t_{-2}	-48.92	-21.70	-57.06
Sent t_{-3}	-10.42	-5.57	-4.44
Sent t_{-4}	-58.66	*-78.42*	-56.58
Sent t_{-5}	-31.51	-8.62	-20.75
VIX t_{-1}	***-107.14****	***-105.35****	***-106.01****
VIX t_{-2}	***-68.60****	**-66.07****	***-68.68****
VIX t_{-3}	*-54.70**	**-54.15****	**-54.30****
VIX t_{-4}	-34.64	*-35.37**	-34.58
VIX t_{-5}	**-50.84****	**-51.68****	**-50.82****
R^2	3.70%	3.88%	3.84%
AIC	-3720.266	-3723.162	-3722.546
$P-Value(Sent)$	*0.084**	**0.05****	**0.028****

Appendix E

Reduced Lag Lexica Performance

The following tables present the results of OLS regression results cross all dictionaries using two feature selection methods genetic algorithms and particle swarm optimization against returns for each of the three indices. Sentiment is calculated as absolute term frequency obtained from the corpus of no exact duplicates. Regression models include a lagged sentiment values from t - 1 to t - 2, similarly lagged values of market returns are also included in the regression model. Each coefficient measures the impact of an increase in one standard deviation on returns in basis points (one basis point equals a return of 0.01%). The regression is based on 1,531 observations, I use Newey and West standard errors similar to Tetlock so as to account for autocorrelation. Italics denotes a significance level of 90% (*), bold denote a significance level of 95% (**) bold and italics denotes a significance level of 99% (***)

Table E.1: Negative DJIA Reduced Lag Refined Lexica

	Negative	$\text{Negative}_{Genetics}$	$\text{Negative}_{Particle}$
Sent t_{-1}	***-26.82****	***-28.96****	***-29.51****
Sent t_{-2}	**18.28****	**18.65****	**20.73****
Dow t_{-1}	***-19.32****	***-19.44****	***-19.24****
Dow t_{-2}	*-12.17**	**-12.08****	*-11.94**
R^2	3.72%	4.22%	4.06%
AIC	-8734.282	-8742.251	-8739.632
$P-Value(Sent)$	***0.008****	***0.006****	***0.003****

Table E.2: Negative-Inf DJIA Reduced Lag Refined Lexica

	Negative-Inf	Negative-Inf $_{Genetics}$	Negative-Inf $_{Particle}$
Sent t_{-1}	***-26.66****	***-28.50****	***-28.75****
Sent t_{-1}	**18.29****	**19.91****	**20.37****
Dow t_{-1}	***-19.52****	***-19.59****	***-19.72****
Dow t_{-2}	*-12.36**	*-11.95**	*-12.45**
R^2	3.74%	4.21%	4.00%
AIC	-8734.5	-8741.9	-8738.6
$P-Value(Sent)$	***0.01****	***0.002****	***0.01****

Table E.3: Positive DJIA Reduced Lag Refined Lexica

	Positive	Positive $_{Genetics}$	Positive $_{Particle}$
Sent t_{-1}	**-23.46****	**-21.77****	***-23.31****
Sent t_{-1}	*15.62**	*13.34**	*15.00**
Dow t_{-1}	**-18.32****	**-18.19****	**-18.25****
Dow t_{-2}	-11.40	-10.93	-11.21
R^2	3.28%	3.24%	3.30%
AIC	-8727.29	-8726.691	-8727.585
$P-Value(Sent)$	*0.07**	*0.06**	**0.033****

Table E.4: Financial Negative DJIA Reduced Lag Refined Lexica

	Fin Neg	Fin Neg $_{Genetics}$	Fin Neg $_{Particle}$
Sent t_{-1}	***-23.03****	***-27.49****	***-23.75****
Sent t_{-2}	*13.84**	**14.81****	12.40
Dow t_{-1}	***-19.78****	***-19.50****	***-19.69****
Dow t_{-2}	*-12.58**	*-12.32**	*-12.62**
R^2	3.45%	4.23%	3.66%
AIC	-8730.039	-8742.458	-8733.328
$P-Value(Sent)$	**0.025****	***0.005****	***0.01****

Table E.5: Negative S&P 500 Reduced Lag Refined Lexica

	Negative	Negative $_{Genetics}$	Negative $_{Particle}$
Sent t_{-1}	***-27.91****	***-30.78****	***-30.53****
Sent t_{-2}	**17.92****	**19.97****	**20.14****
S&P 500 t_{-1}	***-21.50****	***-21.23****	***-21.45****
S&P 500 t_{-2}	*-13.69**	*-13.57**	*-13.8**
R^2	3.64%	4.20%	3.95%
AIC	-8449.096	-8457.995	-8454.008
$P-Value(Sent)$	**0.012****	***0.002****	**0.004****

Table E.6: Negative-Inf S&P Reduced Lag Refined Lexica

	Negative-Inf	Negative-Inf $_{Genetics}$	Negative-Inf $_{Particle}$
Sent t_{-1}	***-27.94***	***-26.43***	***-29.93***
Sent t_{-2}	**18.07**	*16.97*	**19.80**
S&P 500 t_{-1}	***-21.71***	***-21.68***	**-21.63**
S&P 500 t_{-2}	*-13.91*	*-13.74*	*-13.85*
R^2	3.67%	3.60%	3.94%
AIC	-8449.64	-8448.50	-8453.84
$P-Value(Sent)$	***0.01****	**0.025**	***0.008****

Table E.7: Positive S&P 500 Reduced Lag Refined Lexica

	Positive	Positive $_{Genetics}$	Positive $_{Particle}$
Sent t_{-1}	**-24.08****	*-16.69**	**-22.85****
Sent t_{-2}	14.74	7.87	12.07
S&P 500 t_{-1}	***-20.54*****	***20.55*****	***-20.81*****
S&P 500 t_{-2}	-12.81	-12.54	-12.95
R^2	3.23%	2.83%	3.22%
AIC	-8442.60	-8436.27	-8442.38
$P-Value(Sent)$	*0.087**	0.196	*0.081**

Table E.8: Financial Negative S&P 500 Reduced Lag Refined Lexica

	Fin Neg	Fin Neg $_{Genetics}$	Fin Neg $_{Particle}$
Sent t_{-1}	***-24.27*****	***-25.57*****	***-28.63*****
Sent t_{-2}	13.18	13.25	**17.60****
S&P 500 t_{-1}	***-21.99*****	***-21.42*****	***-21.80*****
S&P 500 t_{-2}	*-14.16**	-13.87	*-13.98**
R^2	3.45%	3.67%	3.86%
AIC	-8446.032	-8449.636	-8452.543
$P-Value(Sent)$	**0.031****	**0.016****	***0.006*****

Table E.9: Negative VIX Reduced Lag Refined Lexica

	Negative	Negative $_{Genetics}$	Negative $_{Particle}$
Sent t_{-1}	**73.09****	**60.41****	**71.02****
Sent t_{-2}	**-61.73****	**-48.12****	*-58.02**
VIX t_{-1}	***-96.95*****	***96.34*****	***-97.33*****
VIX t_{-2}	**-53.02***	*-52.44**	*-51.93**
R^2	2.52%	2.42%	2.48%
AIC	-3723.646	-3722.101	-3723.105
$P-Value(Sent)$	*0.088**	*0.072**	0.115

Table E.10: Negative-Inf VIX Reduced Lag Refined Lexica

	Negative-Inf	Negative-Inf $_{Genetics}$	Negative-Inf $_{Particle}$
Sent t_{-1}	**73.32****	**65.6****	**79.1****
Sent t_{-2}	**-61.61****	*-52.44**	**-65.94****
VIX t_{-1}	***-97.38*****	***-96.70*****	***-96.62*****
VIX t_{-2}	**-53.37****	*-51.57**	*-53.01**
R^2	2.53%	2.47%	2.62%
AIC	-3723.914	-3722.909	-3725.206
$P-Value(Sent)$	*0.0973**	0.12	*0.065**

Table E.11: Positive VIX Reduced Lag Refined Lexica

	Positive	Positive $_{Genetics}$	Positive $_{Particle}$
Sent t_{-1}	54.83	46.77	57.57
Sent t_{-2}	-39.95	-33.31	-41.82
VIX t_{-1}	***-95.75*****	***-95.68*****	***-95.58*****
VIX t_{-2}	*-51.02**	*-50.28**	*-51.08**
R^2	2.30%	2.26%	2.33%
AIC	-3720.2	-3719.5	-3720.6
$P-Value(Sent)$	0.357	0.319	0.3

Table E.12: Financial Negative VIX Reduced Lag Refined Lexica

	Fin Neg	Fin Neg $_{Genetics}$	Fin Neg $_{Particle}$
Sent t_{-1}	**67.75****	**67.84****	**77.55****
Sent t_{-2}	*-54.83**	-42.86	**-65.56****
VIX t_{-1}	***-98.12*****	***-96.72*****	***-97.24*****
VIX t_{-2}	**-53.63****	*-52.50**	**-53.84****
R^2	2.50%	2.52%	2.65%
AIC	-3723.318	-3723.755	-3725.75
$P-Value(Sent)$	0.115	*0.0973**	**0.047****

Bibliography

[1] ABBASI, A., CHEN, H., AND SALEM, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst. 26*, 3 (June 2008), 12:1–12:34.

[2] AHMAD, K. Edderkoppspinn eller nettverk: News media and the use of polar words in emotive contexts, 2008.

[3] AHMAD, K., DALY, N., AND LISTON, V. What is new? news media, general elections, sentiment, and named entities. *Sentiment Analysis where AI meets Psychology (SAAIP)* (2011), 80.

[4] AHMAD, K., GILLAM, L., AND TOSTEVIN, L., E. A. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text Retrieval Conference (TREC-8)* (1999).

[5] ANTWEILER, W., AND FRANK, M. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance 59*, 3 (2004), 1259–1294.

[6] ATKINSON-ABUTRIDY, J., MELLISH, C., AND AITKEN, S. Combining information extraction with genetic algorithms for text mining. *Intelligent Systems, IEEE 19*, 3 (2004), 22–30.

[7] BÄECK, T., HOFFMEISTER, F., AND SCHWEFEL, H.-P. A survey of evolution strategies. In *Proceedings of the Fourth International Conference on Genetic Algorithms* (1991), Morgan Kaufmann, pp. 2–9.

[8] BLICKLE, T., AND THIELE, L. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation 4*, 4 (1996), 361.

[9] BOLLEN, J., MAO, H., AND ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1–8.

[10] BOULIS, C., AND OSTENDORF, M. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. In *Proc. of the International Workshop in Feature Selection in Data Mining* (2005), pp. 9–16.

[11] BRANDENBURG, H. Political bias in the irish media: A quantitative study of campaign coverage during the 2002 general election. *Irish Political Studies 20*, 3 (2005), 297–322.

[12] CAMPBELL, J., GROSSMAN, S., AND WANG, J. Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics 108*, 4 (1993), 905–939.

[13] CLERC, M., AND KENNEDY, J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on 6*, 1 (2002), 58–73.

[14] COBB, H. G. Genetic algorithms for tracking changing environments. In *Proceedings of the Fifth International Conference on Genetic Algorithms* (1993), Morgan Kaufmann, pp. 523–530.

[15] COLE, N., LOUIS, S., AND MILES, C. Using a genetic algorithm to tune first-person shooter bots. In *Evolutionary Computation, 2004. CEC2004. Congress on* (june 2004), vol. 1, pp. 139 – 145 Vol.1.

[16] DA, Z., ENGELBERG, J., AND GAO, P. In search of attention. *The Journal of Finance 66*, 5 (2011), 1461–1499.

[17] DA, Z., ENGELBERG, J., AND GAO, P. The sum of all fears: Investor sentiment and asset prices, 2011.

[18] DALY, N., AHMAD, K., AND KEARNEY, C. Correlating market movements with consumer confidence and sentiments: A longitudinal study. *Text Mining Services 2009* (2009), 169–180.

[19] DAS, D., AND BANDYOPADHYAY, S. Sentence-level emotion and valence tagging. *Cognitive Computation* (2012), 1–16.

[20] DE JONG, K., AND SPEARS, W. E. A. On the state of evolutionary computation. In *Proceedings of the Fifth International Conference on Genetic Algorithms* (1993), Citeseer, pp. 618–626.

[21] DEVITT, A. *Methods for Meaninful Text Representation and Comparison*. Ph.D. Thesis, Computational Linguistic Group, Trinity College Dublin, 2005.

[22] DOUGAL, C., ENGELBERG, J., GARCÍA, D., AND PARSONS, C. Journalists and the stock market. *Review of Financial Studies 25*, 3 (2012), 639–679.

[23] EBERHART, R., AND KENNEDY, J. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS '95., Proceedings of the Sixth International Symposium on* (oct 1995), pp. 39 –43.

[24] EKRT, A., AND NMETH, S. Maintaining the diversity of genetic programs. In *Genetic Programming*, J. Foster, E. Lutton, J. Miller, C. Ryan, and A. Tettamanzi, Eds., vol. 2278 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2002, pp. 122–135.

[25] ESHELMAN, L., AND SCHAFFER, J. Preventing premature convergence in genetic algorithms by preventing incest. In *Proceedings of the Fourth International Conference on Genetic Algorithms* (1991), vol. 115, p. 122.

[26] FAN, H.-Y., LU, J. W.-Z., AND XU, Z.-B. An empirical comparison of three novel genetic algorithms. *Engineering Computations 17*, 8 (2000), 981–1002.

[27] FOGEL, D. Evolutionary computation: A new transactions. *Evolutionary Computation, IEEE Transactions on 1*, 1 (apr 1997), 1 –2.

[28] FOGEL, D., AND ATMAR, J. Comparing genetic operators with gaussian mutations in simulated evolutionary processes using linear systems. *Biological Cybernetics 63* (1990), 111–114.

[29] FOLI, K., OKABE, T., OLHOFER, M., JIN, Y., AND SENDHOFF, B. Optimization of micro heat exchanger: Cfd, analytical approach and multi-objective evolutionary algorithms. *International Journal of Heat and Mass Transfer 49*, 56 (2006), 1090 – 1099.

[30] GARCIA, D. Sentiment during recessions. *Journal of Finance, Forthcoming* (2012).

[31] GARCIA, D., GARAS, A., AND SCHWEITZER, F. Positive words carry less information than negative words. *arXiv preprint arXiv:1110.4123* (2011).

[32] GEHLHAAR, D. K., VERKHIVKER, G. M., REJTO, P. A., SHERMAN, C. J., FOGEL, D. R., FOGEL, L. J., AND FREER, S. T. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & amp; Biology 2*, 5 (1995), 317 – 324.

[33] GILLAM, L., AND AHMAD, K. Pattern mining across domain-specific text collections. *Machine Learning and Data Mining in Pattern Recognition* (2005), 634–634.

[34] GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[35] GREFENSTETTE, J. Optimization of control parameters for genetic algorithms. *Systems, Man and Cybernetics, IEEE Transactions on 16*, 1 (jan. 1986), 122 –128.

[36] HAUPT, R. Optimum population size and mutation rate for a simple real genetic algorithm that optimizes array factors. In *Antennas and Propagation Society International Symposium, 2000. IEEE* (2000), vol. 2, pp. 1034 –1037 vol.2.

[37] HENRY, E., AND LEONE, A. Measuring qualitative information in capital markets research. *Available at SSRN 1470807* (2009).

[38] HO, S., LIN, H., LIAUH, W., AND HO, S. Opso: Orthogonal particle swarm optimization and its application to task assignment problems. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 38*, 2 (2008), 288–298.

[39] HOLLAND, J. H. *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.

[40] HOLSTI, O. *Content analysis for the social sciences and humanities*. Addison-Wesley Reading, MA, 1969.

[41] HORN, J., NAFPLIOTIS, N., AND GOLDBERG, D. E. A niched pareto genetic algorithm for multiobjective optimization. In *IEEE World Congress on Computational Intelligence* (1994).

[42] HU, M., AND LIU, B. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2004), KDD '04, ACM, pp. 168–177.

[43] HUTH, A., AND WISSEL, C. The simulation of fish schools in comparison with experimental data. *Ecological Modelling 7576*, 0 (1994), 135 – 146.

[44] KAHNEMAN, D., AND TVERSKY, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* (1979), 263–291.

[45] KENNEDY, J. The particle swarm: social adaptation of knowledge. In *Evolutionary Computation, 1997., IEEE International Conference on* (apr 1997), pp. 303 –308.

[46] KENNEDY, J., AND EBERHART, R. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on* (nov/dec 1995), vol. 4, pp. 1942 –1948 vol.4.

[47] KENNEDY, J., AND EBERHART, R. A discrete binary version of the particle swarm algorithm. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on* (oct 1997), vol. 5, pp. 4104 –4108 vol.5.

[48] KHALIESSIZADEH, S., ZAEFARIAN, R., NASSERI, S., AND ARDIL, E. Genetic mining: using genetic algorithm for topic based on concept distribution. In *Proceedings of World Academy of Science, Engineering and Technology* (2006), vol. 13, Citeseer, pp. 1307–8884.

[49] KILGARRIFF, A., AND ROSE, T. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing* (1998), Citeseer, pp. 46–52.

[50] KIM, S.-M., PANTEL, P., CHKLOVSKI, T., AND PENNACCHIOTTI, M. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2006), EMNLP '06, Association for Computational Linguistics, pp. 423–430.

[51] KOPPEL, M., AND SHTRIMBERG, I. Good news or bad news let the market decide. In *Computing Attitude and Affect in Text Theory and Applications*, J. Shanahan, Y. Qu, and J. Wiebe, Eds., vol. 20 of *The Information Retrieval Series*. Springer Netherlands, 2006, pp. 297–301.

[52] KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Mit Press, 1992.

[53] KRIPPENDORFF, K. *Content analysis: An introduction to its methodology*. Sage Publications, Incorporated, 2012.

[54] LASSWELL, H. *The comparative study of symbols: An introduction*. Stanford University Press, 1952.

[55] LASSWELL, H. E. A. *Power and personality*. Transaction Pub, 2009.

[56] LAVER, M., AND BENOIT, K. Locating tds in policy spaces: The computational text analysis of dail speeches. *Irish Political Studies 17*, 1 (2002), 59–73.

[57] LAVER, M., BENOIT, K., AND GARRY, J. Extracting policy positions from political texts using words as data. *American Political Science Review 97*, 02 (2003), 311–331.

[58] LAVER, M., AND GARRY, J. Estimating policy positions from political texts. *American Journal of Political Science* (2000), 619–634.

[59] LEOPOLD, E., AND KINDERMANN, J. Text categorization with support vector machines. how to represent texts in input space. *Machine Learning* (2002).

[60] LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. 1966 (8). *Forschungsbericht.–707–710 S* (1966).

[61] LI, X., AND YAO, Y. Cooperatively coevolving particle swarms for large scale optimization. *IEEE Transactions on Evolutionary Computation 16*, 2 (2011), 1–15.

[62] LIN, J., AND YU, J. Weighted naive bayes classification algorithm based on particle swarm optimization. In *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on* (2011), IEEE, pp. 444–447.

[63] LIU, Y., YAO, X., ZHAO, Q., AND HIGUCHI, T. Scaling up fast evolutionary programming with cooperative coevolution. In *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on* (2001), vol. 2, pp. 1101 –1108 vol. 2.

[64] LOUGHRAN, T., AND MCDONALD, B. When is a liability not a liability textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*, 1 (2011), 35–65.

[65] LOUIS, S., AND MILES, C. Playing to learn: case-injected genetic algorithms for learning to play computer games. *Evolutionary Computation, IEEE Transactions on 9*, 6 (dec. 2005), 669 – 681.

[66] MAN, K. F., TANG, K. S., AND KWONG, S. *Genetic Algorithms: Concepts and Designs*, 2nd ed. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999.

[67] MANDELBROT, B., HUDSON, R., AND GRUNWALD, E. The (mis)behaviour of markets. *The Mathematical Intelligencer 27* (2005), 77–79.

[68] MARSHALL, B. R., VISALTANACHOTI, N., AND COOPER, G. Sell the rumour, buy the fact? *Accounting & Finance* (2012), no–no.

[69] MARTINEAU, J., AND FININ, T. Delta tfidf:an improved feature space for sentiment analysis. In *Proceedings of the Third AAAI Internatonal Conference onWeblogs and SocialMedia* (2009), AAAI Press.

[70] MATHE, E., AND GREFENSTETTE, J. Polyoptimizing genetic algorithm for feature subset selection. In *Interface 2004: Classification and Clustering 36th Symposium on the Interface* (2004).

[71] MOSER, A., AND NARASIMHA MURTY, M. On the Scalability of Genetic Algorithms to Very Large-Scale Feature Selection. In *Real-World Applications of Evolutionary Computing*, vol. 1803 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2000, pp. 309–311.

[72] MUKHERJEE, I., AL-FAYOUMI, M., MAHANTI, P., JHA, R., AND AL-BIDEWI, I. Content analysis based on text mining using genetic algorithm. In *Computer Technology and Development (ICCTD), 2010 2nd International Conference on* (2010), pp. 432–436.

[73] MUKRAS, R., WIRATUNGA, N., LOTHIAN, R., CHAKRABORTI, S., AND HARPER, D. Information gain feature selection for ordinal text classification using probability re-distribution. In *Proceedings of the Textlink workshop at IJCAI* (2007), vol. 7.

[74] MUYL, F., DUMAS, L., AND HERBERT, V. Hybrid method for aerodynamic shape optimization in automotive industry. *Computers & Fluids 33*, 56 (2004), 849 – 858.

[75] NAMENWIRTH, J., AND LASSWELL, H. *The changing language of American values: a computer study of selected party platforms*. Sage Publications, 1970.

[76] NEUENDORF, K. A. *The content analysis guidebook*. Sage, 2002.

[77] OKEEFE, T., AND KOPRINSKA, I. Feature selection and weighting in sentiment analysis. In *Proceedings of the 14th Australian Document Computing Symposium* (2009).

[78] OKEEFE, T., AND KOPRINSKA, I. Feature selection and weighting methods in sentiment analysis. *ADCS 2009* (2009), 67.

[79] ONEILL, M., BRABAZON, A., RYAN, C., AND COLLINS, J. Evolving market index trading rules using grammatical evolution. In *Applications of Evolutionary Computing*, E. Boers, Ed., vol. 2037 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2001, pp. 343–352.

[80] PALTOGLOU, G., AND THELWALL, M. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), ACL '10, Association for Computational Linguistics, pp. 1386–1395.

[81] PANG, B., LEE, L., AND VAITHYANATHAN, S. Thumbs up? sentiment analysis classification using machine learning techniques. In *In Proceedings of the Conference Conference on Empirical Methods on Natural Language Processing* (2002).

[82] PIETRAMALA, A., POLICICCHIO, V. L., RULLO, P., AND SIDHU, I. A genetic algorithm for text classification rule induction. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 188–203.

[83] QUIRK, R., AND CRYSTAL, D. *A comprehensive grammar of the English language*, vol. 6. Cambridge Univ Press, 1985.

[84] REYNOLDS, C. W. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1987), SIGGRAPH '87, ACM, pp. 25–34.

[85] ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation 60*, 5 (2004), 503–520.

[86] ROMEY, W. L. Individual differences make a difference in the trajectories of simulated schools of fish. *Ecological Modelling 92*, 1 (1996), 65 – 77.

[87] ROYALL, R. *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, 1997.

[88] RUIZ, E., HRISTIDIS, V., CASTILLO, C., GIONIS, A., AND JAIMES, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), ACM, pp. 513–522.

[89] SCHILLER, R. *The Irrational Exuberance*. Wiley Online Library, 2000.

[90] SCHUMAKER, R., AND CHEN, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS) 27*, 2 (2009), 12.

[91] SCHWEFEL, H.-P. *Numerical Optimization of Computer Models*. John Wiley & Sons, Inc., New York, NY, USA, 1981.

[92] SHAFFER, J., AND ESHELMAN, L. J. On crossover as an evolutionarily viable strategy. In *Proceesing of the Fourth International Conference on Genetic Algorithms* (1991), pp. 61–68.

[93] SHI, Y., AND EBERHART, R. Empirical study of particle swarm optimization. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on* (1999), vol. 3, IEEE.

[94] SILVERMAN, D. *Qualitative research*. Sage Publications Limited, 2010.

[95] SONODA, T., YAMAGUCHI, Y., ARIMA, T., OLHOFER, M., SENDHOFF, B., AND SCHREIBER, H.-A. Advanced high turning compressor airfoils for low reynolds number condition—part i: Design and optimization. *Journal of Turbomachinery 126*, 3 (2004), 350–359.

[96] SOUKOREFF, R., AND MACKENZIE, I. Measuring errors in text entry tasks: An application of the levenshtein string distance statistic. In *CHI'01 extended abstracts on Human factors in computing systems* (2001), ACM, pp. 319–320.

[97] SPEARS, W. M. Crossover or mutation. In *Foundations of Genetic Algorithms 2* (1993), Morgan Kaufmann, pp. 221–237.

[98] SPEARS, W. M., AND JONG, K. A. D. On the virtues of parameterized uniform crossover. In *In Proceedings of the Fourth International Conference on Genetic Algorithms* (1991), pp. 230–236.

[99] SRINIVAS, M., AND PATNAIK, L. Adaptive probabilities of crossover and mutation in genetic algorithms. *Systems, Man and Cybernetics, IEEE Transactions on 24*, 4 (1994), 656–667.

[100] STONE, P. J., DUNPHY, D. C., AND SMITH, M. S. *THE GENERAL INQUIRER: A COMPUTER APPROACH TO CONTENT ANALYSIS*. MIT Press, 1966.

[101] TAYLOR, S. *Asset price dynamics, volatility, and prediction*. Princeton university press, 2011.

[102] TEICHERT, T., HEYER, G., SCHÖNTAG, K., AND MAIRIF, P. Co-word analysis for assessing consumer associations: a case study in market research. In *Affective Computing and Sentiment Analysis*. Springer, 2011, pp. 115–124.

[103] TETLOCK, P., SAAR-TSECHANSKY, M., AND MACSKASSY, S. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance 63*, 3 (2008), 1437–1467.

[104] TETLOCK, P. C. Giving content to investor sentiment the role of media in the stock market. *The Journal of Finance 62*, 3 (2007), 1139–1168.

[105] TETLOCK, P. C. All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies 24*, 5 (2011), 1481–1512.

[106] THOMAS, J., AND SYCARA, K. Integrating genetic algorithms and text learning for financial prediction. In *In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*. (2000).

[107] TONG, R. M. An operational system for detecting and tracking opinions in on-line discussion. In *Workshop on Operational Text Classication.* (2001), SIGIR.

[108] TRELEA, I. The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information processing letters 85*, 6 (2003), 317–325.

[109] TU, X., AND TERZOPOULOS, D. Artificial fishes: physics, locomotion, perception, behavior. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1994), SIGGRAPH '94, ACM, pp. 43–50.

[110] TURNEY, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 2002), ACL 2002, Association for Computational Linguistics, pp. 417–424.

[111] VATTER, H., AND WALKER, J. *History of the US Economy Since World War II.* ME Sharpe Inc, 1996.

[112] YANG, J., AND HONAVAR, V. Feature subset selection using a genetic algorithm. *Intelligent Systems and their Applications, IEEE 13*, 2 (mar/apr 1998), 44 –49.

[113] YANG, Y., AND PEDERSEN, J. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* (1997), MORGAN KAUFMANN PUBLISHERS, INC., pp. 412–420.

[114] YANG, Z., TANG, K., AND YAO, X. Differential evolution for high-dimensional function optimization. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on* (2007), IEEE, pp. 3523–3530.

[115] YANG, Z., TANG, K., AND YAO, X. Multilevel cooperative coevolution for large scale optimization. In *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on* (june 2008), pp. 1663 –1670.

[116] YENIAY, O. Penalty function methods for constrained optimization with genetic algorithms. *Mathematical and Computational Applications 10*, 1 (2005), 45–56.

[117] ZIRN, C., NIEPERT, M., STUCKENSCHMIDT, H., AND STRUBE, M. Fine-grained sentiment analysis with structural features. In *IJCNLP* (2011), pp. 336–344.