Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Is There a Language of Sentiment?

An Analysis of Lexical Resources for Sentiment Analysis

Ann Devitt · Khurshid Ahmad

February 5, 2013

Abstract In recent years, sentiment analysis (SA) has emerged as a rapidly expanding field of application and research in the area of information retrieval. In order to facilitate the task of selecting lexical resources for automated SA systems, this paper sets out a detailed analysis of four widely used sentiment lexica. The analysis provides an overview of the coverage of each lexicon individually, the overlap and consistency of the four resources and a corpus analysis of the distribution of the resources' lexical contents in general and specialised language. This work aims to explore the characteristics of affective language as represented by these lexica and the implications of the findings for developers of sentiment analysis systems.

Keywords Sentiment analysis · electronic lexica · corpus analysis · Financial Information Extraction

1 Introduction

In recent years, the area of sentiment analysis (SA) in text has become a focus of attention in the fields of theoretical and computational linguistics, investigating the production and processing of affective contours in text, the textual corollary of emotional prosody in speech. Extensive research has been conducted in the area developing automated sentiment analysis systems and resources to build and test these systems, such as annotated corpora or lexica.

A. Devitt
Trinity College Dublin, Ireland.
Tel: 353 1 896 1293
Fax: 353 1 677 7238
E-mail: ann.devitt@tcd.ie

K. Ahmad
Trinity College Dublin, Ireland.

Research has drawn on text from many domains ranging from on-line film reviews (Turney 2002) to newspaper editorials (Wiebe et al. 2003) to Dow Jones News Service headlines (Mitchell and Mulherin 1994). While there is much work done which uses polarity or sentiment lexica and some work done to derive sentiment lexica, to the authors' knowledge, this is the first comparative analysis of the contents and characteristics of the lexical resources available for sentiment analysis and what are the distributional characteristics of these terms and features in general language.

In selecting a resource to aid analysis of emotion in language, it is necessary to evaluate the potential contributions or drawbacks that resource offers relative to others. This paper sets out to examine four sentiment lexica used in the Sentiment Analysis literature (the General Inquirer lexicon (GI) (Stone et al. 1966), Dictionary of Affect in Language (DAL) (Whissell 1989), SentiWordNet (SWN) version 1.0 (Esuli and Sebastiani 2006) and WordNet-Affect (WNA) (Strappavara and Valitutti 2004)) to define the cognitive and empirical bases and the emotional spectrum or bias of what to date have been considered the lexica of emotion. These lexica were selected to represent a cross-section of widely used resources across the range of sentiment analysis applications from movie reviews to financial news and to provide a range of modes of development, theoretical and disciplinary backgrounds and types of annotation. The structure, content and encoding of the lexica as well as their relative distributions in general language and sub-languages of English are analysed to determine the characteristics of the language encoded in sentiment lexica and whether there is sufficient consistency of content and coverage between the lexica to claim that the lexica provide an albeit limited but coherent representation of the language of emotion as it is used in English. Furthermore, the implications of the findings are drawn out to emphasise the possible contributions as well as possible bias or error introduced by the lexica used alone or in combination in SA systems.

The research context for this investigation is set out in section 2. Current cognitive theories of what constitutes emotion underpin this investigation of the language of emotion and are outlined in section 2.1. Section 2.2 outlines approaches to sentiment analysis in the literature and in particular the role of prior polarity as defined in sentiment lexica in sentiment analysis applications. Section 3 provides a thorough analysis of the four lexica in terms of structure and content individually and in relation to each other. Given that the four sentiment lexica differ in many ways, as will be outlined in section 3.1, the task of selecting a sentiment lexicon for a sentiment analysis application in a given domain can be an onerous task. To facilitate such a task, section 3.2 provides a comparative evaluation of the lexical content of the resources and section 3.3 compares them in terms of the manner and consistency of sentiment representation. The aim is to establish the degree of consensus between resources in both respects. Strong consensus provides a means of validating the resources as reliable repositories of consistent information about emotion or sentiment in language. An analysis of consensus also establishes if the re-

sources are mutually exclusive and can provide guidelines for using lexica alone or in combination.

In order to open a discussion on what the lexica of sentiment represent, section 4 sets out a corpus analysis of the use of sentiment lexicon terms in general language and specific genres or domains. The aim of the general language corpus analysis in section 4.2 is to determine whether the terms encoded in the sentiment lexicon share characteristics and patterns of use that distinguish them from an arbitrary collection of lexical items. More precisely, do these resources represent a coherent set of affective lexical items that share distributional features in language which make them distinctive from "general language"? The analysis would suggest that, yes, the sentiment lexica do encode a coherent lexicon of emotion that functions in a particular fashion. A secondary aim is to determine the usefulness of each lexical resource in terms of the distribution of their lexical items and features, in particular polarity features, in general language. This analysis is invaluable for those working in automatic sentiment analysis to determine the coverage and orientation of available resources and potentially specify requirements for new or extended lexical resources. The further comparative corpus analysis set out in section 4.3 aims to determine whether lexical and polarity distributions differ across varieties or sub-languages of English and in what ways. This more focused analysis highlights the domain–dependent nature of affect and polarity in text and the possible need to re–assess resource requirements and underlying assumptions for sentiment analysis applications in different domains.

2 Research Context

2.1 Current Psychological Theories of Emotion

In order to understand how emotion can be realised in text, we must first have a notion of what emotion is and how people experience it. Current cognitive theories of what constitutes emotion underpin this investigation of the language of emotion. There are two primary approaches to a cognitive account of emotion:

- emotion as finite categories;
- emotion as dimensions.

Computational linguistics has largely espoused the dimensional model of emotion. This section sets out the on-going debate in psychology regarding an accurate model of emotion and emotional experience. The categorical approach posits a finite set of basic emotions which are experienced universally across cultures. The basic emotion set posited by researchers can vary according to different accounts and cultural contexts but generally include happiness, sadness, anger, disgust and fear (Ekman and Friesen 1971). The theory is strongly supported by evidence for categorical perception of facial expressions (Etcoff and Magee 1992). However, the results in word perception tasks are less conclusive (Niedenthal and Halberstadt 2000). The dimensional approach delineates

emotions not according to discrete categories but rather multiple dimensions on which all emotional states, emotional dispositions or affective appraisals can be plotted. Russell and Mehabrian (1977) and Osgood et al. (1957) for example distinguish three dimensions:

- good–bad axis (termed the dimension of valence, evaluation or pleasantness)
- active–passive axis (termed the dimension of arousal, activation or intensity)
- strong–weak axis (termed the dimension of dominance or submissiveness)

The two primary dimensions in the literature and those found consistently across a series of emotion dimension experiments by Watson and Tellegen (1985) are valence and arousal. This two dimensional model of emotion is illustrated in figure 1. Russell (1980) also posits a unified or "circumplex" model combining these dimensions whereby the emotional space can be represented as a circle where any emotion can be located on this bidimensional plane relative to its two axes of valence and arousal. The debate between the categorical and dimensional approaches is on–going, as evidenced by the 1994 dispute of Russell and Ekman published in the Psychological Bulletin (Russell 1994; Ekman 1994). It is likely that a unified account will be required whereby some emotions are categorical and "pre-wired" while others are dimensional, not innate and based on higher-level processes. Whatever the theory of emotion one chooses to espouse, there is strong evidence that emotion or mood impacts on other cognitive processes such as memory and decision–making and this motivates much work in fields as diverse as behavioural finance, neuroscience and linguistics.

The categorical–dimensional debate has an impact on a computational approach to emotion as expressed in language and on any linguistic resources used in that approach. The categorical approach posits a finite set of *primary* emotions. However, this set of emotions is not exhaustive and does not cover all emotionally–charged experience or indeed text but rather a subset of discrete non–decomposable emotional states. Other emotional experience which cannot be categorised as one of the primary emotions could be said to be secondary but no theory provides an exhaustive categorisation, hierarchical or otherwise, of all human emotional experience. For this reason, the dimensional theory is perhaps more amenable to the representation of emotional experience in general, as any experience can be located somewhere in a multi-dimensional emotional space, not just at fixed points in that space. As the field of computational sentiment analysis aims to evaluate free text on any topic rather than representations of prototypical emotions (such as facial expressions), a dimensional representation of emotion is appropriate here and allows enough flexibility to estimate degrees and shades of emotion. This is reflected in the uptake of the dimensional representation of emotion for the construction and elaboration of all the emotion lexica discussed below. Furthermore, the strong focus on valence or the positive–negative dimension of emotion in sentiment analysis and its resources is justified to a certain degree by strong evidence in

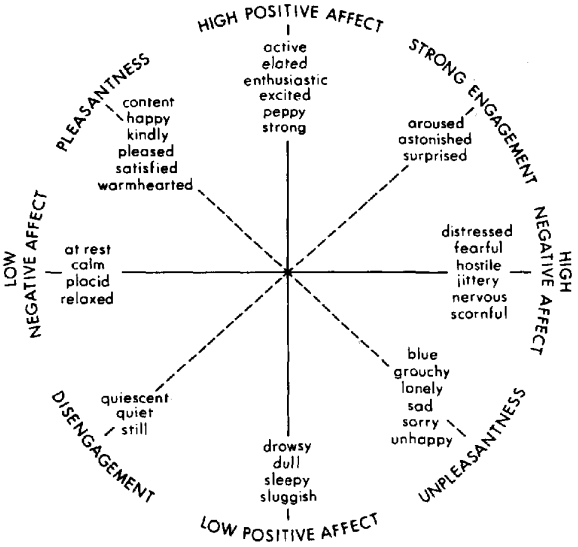


Fig. 1 Two factor structure of affect (Watson and Tellegen 1985, p.22)

the psychology literature for valence having the greatest impact on cognitive processes relative to other emotion dimensions (Niedenthal and Halberstadt 2000, p.173). While the modelling of emotion in psychology remains an open question, the espousal of the dimensional model in computational linguistics, as evidenced by both polarity identification systems and their resources, is both practical and justified.

2.2 Current Approaches to Sentiment Analysis

Since the 1990s, Sentiment Analysis (SA) has emerged as a field of intensive research in information retrieval and computational linguistics. The term covers a range of tasks related to the automatic identification of aspects of affective content in unimodal input, such as text or multimodal input, such as video. The tasks range from word– to document–level analysis, coarse-grained identification of subjectivity to fine-grained attribution of specific opinions, single to multiple domain input across a variety of languages. Many current approaches use machine learning (ML) techniques to build affective text classifiers from data, tagged (supervised ML) or untagged (unsupervised ML) using a variety of algorithms (Naive Bayes, Support Vector Machines, Bayesian Belief Networks, etc) (Kim and Hovy 2004; Wiebe et al. 2004; Hatzivassiloglou and Wiebe 2000). In recent years the vast increases in memory and processing speeds as well as new developments in machine learning algorithms have led to vast improvements in ML results for many NLP tasks, sentiment analysis

included. A key factor in determining the success of a ML approach is the quality and volume of the training data. The more data and the more consistent and noise-free it is, the better the results. Recent years have seen an increase in the compilation and release of sentiment analysis data resources for machine learning. On-line product reviews have provided a rich source of data with customer ratings interpreted as document–level sentiment orientation and intensity ratings (Pang et al. 2002; Blitzer et al. 2007). Blogs with writer mood ratings have been used in a similar fashion (Mihalcea and Liu 2006). The MPQA opinion annotated corpus provides a more detailed human-annotated resource with word– and phrase–level tags for a variety of opinion types (Wiebe et al. 2005). However, it is not data alone which determines ML success, feature selection is also an important factor. Many approaches restrict themselves to using presence or frequency of n-grams of tagged text data with greater success than simple word-counting approaches (Pang et al. 2002) but others use additional linguistic features such as part-of-speech, term position in the text or presence of negation to improve performance (Pang and Lee 2004; Wilson et al. 2005). The accuracy of machine learning models improves year on year with the introduction of new algorithms, features and feature selection mechanisms.

Other approaches rely rather on explicit manipulation of linguistic features which have been identified within a theoretical framework, from introspection or through corpus analysis: Kanayama et al. (2004), for example, adapt a machine translation transfer engine to output sentiment units based on predefined lexical items and sentiment patterns; Kennedy and Inkpen (2006) exploit contextual valence shifters (Polanyi and Zaenen 2004) in an affective lexical item frequency-based implementation; Ahmad et al. (2006) use corpus-derived sentiment regular expression to identify polarity of financial news; Nasukawa and Yi (2003) define a lexicon for transfer of polarity between syntactic arguments; Devitt and Ahmad (2007) apply a theory of text cohesion to weight the contribution of polarity items in text. Although machine learning methods have been successful for sentiment analysis, an analysis of what contributes to the realisation of emotional or affective content in language, building on the work of Polanyi and Zaenen (2004); Bolasco and della Ratta-Rinaldi (2004), for example, is becoming necessary in order both to push the boundaries of performance of existing approaches and to better understand the cognitive processes by which such language is produced and processed. Prior polarity of lexical items is a key linguistic feature for both ML and non-ML approaches and performance can depend on how prior polarity is contextualised, mitigated or intensified by other features in a system. Unigram ML techniques implicitly build a polarity lexicon, some researchers have set out to learn such a lexicon from corpora (Turney 2002; Hatzivassiloglou and McKeown 1997) and many use existing sentiment lexica for implementation (Kennedy and Inkpen 2006; Devitt and Ahmad 2007; Wilson et al. 2005) or evaluation (Turney and Littman 2003; Bolasco and della Ratta-Rinaldi 2004). This paper constitutes a timely contribution in providing an analysis of some of the most widely-used resources in the field.

3 Lexical Resources for Emotion

The domain of sentiment analysis in computational linguistics and information retrieval is quite young but it has the advantage of drawing on long-established work in psychology, linguistics and literature for its theories, resources and evaluation criteria. This section examines a set of four lexical resources available which have been widely used in developing automated sentiment identification techniques:

- General Inquirer (Stone et al. 1966);
- Dictionary of Affect in Language (Whissell 1989);
- WordNet Affect (Strappavara and Valitutti 2004);
- SentiWordNet (Esuli and Sebastiani 2006).

The resources were selected to provide a range of different approaches in terms of the traditions from which they derive, their theoretical underpinning and their representation of emotion or emotional experience.[1] Section 3.1 specifies what each resource claims to represent and how this is encoded in the lexicon. A comparative analysis of lexical content is set out in sections 3.2 and 3.3 to determine to what extent these resources may be complementary, mutually exclusive or indeed contradictory. Section 3.4 outlines the implications of the findings for SA systems.

3.1 The Lexica

Each of the four lexica under analysis derives from quite different theoretical frameworks and the respective underlying assumptions impact on the selection and encoding of terms within each resource. This is realised as differences in development criteria where the lexica rely to different extents on corpus, manual and automatic processes for term selection and sentiment feature identification. These underlying differences in sources and rigour of development could impact on the degree to which the lexicon may be representative of general language, its robustness and accuracy. The differences in development criteria between the lexicon are summarised in table 1. The lexicalisation of emotion is intrinsic to the psychological theories of emotion set out in section 2.1 both as a means of verbalising the theory and as raw material for psychological experiments to study and validate the theory: GI and the DAL derive from this tradition of examining how emotion is realised in text. As table 1 illustrates, both of these lexica rely on corpus analysis to identify salient frequent terms for inclusion in the lexicon. GI supplements the corpus frequent word selection with an additional word list to ensure full coverage. The DAL

[1] An analysis of the overlap of the selected resources with the MPQA dictionary (Wilson et al. 2005), another widely used and freely available resource, was conducted. Over 90% of MPQA overlaps with other resources and of the unique 10% the vast majority of terms are morphological or orthographic variants of shared terms or rare lexical items. The authors deemed the four lexica investigated in depth here sufficiently representative of sentiment lexica to provide comprehensive findings.

validates the corpus selection by hand. Both lexica focus on which terms are used in practice to realise emotion as identified in corpora and by annotators. WordNet Affect is more in the lexicography tradition of domain terminology definition and relies on a hand-coded selection of terms which is then automatically extended. SentiWordNet on the other hand is machine generated, using a very small initial seed set of sentiment terms and relying on automatic classification to determine sentiment polarity ratings. The sections that follow outline the provenance, coverage and contents of each of the lexica in turn.

Table 1 Lexicon Development Criteria

Development Criterion	**Lexica**			
	DAL	**GI**	**WNAffect**	**SentiWN**
Corpus analysis for word selection	Yes	Yes	-	-
Contents validated by hand or corpus	Yes	Yes	-	-
Manual word list	-	Yes	Yes (2000 words)	Yes (20 words)
Automatic expansion of word list	-	-	Yes	Yes

3.1.1 General Inquirer

Provenance. General Inquirer (GI) was developed by Philip Stone at Harvard in the late 1960s (Stone et al. 1966) in the tradition of content analysis and more specifically the lexicalisation of emotion. It is composed of two frequent word lists drawn from two corpora of North American written English at different time periods:

- the Harvard IV dictionary drawn from the Thorndike-Lorge 1920s-1940s corpus (Thorndike and Lorge 1944);
- the Lasswell dictionary from pre-1950 and updated in 1980s: Lasswell and Kaplan (1950); Namenwirth and Weber (1987).

The GI lexicon was validated and tagged by hand according to a broad set of semantic categories motivated by theories in psychology and content analysis.

Representation of Lexical Items and Emotion. The full lexicon contains 8,641 terms, some with multiple senses encoded in the lexicon, tagged for a variety of semantic categories. In total there are 11,788 word senses and 184 possible binary semantic categories relating to domains of use, polarity, social categories, etc. Following an analysis of the approximately thirty sentiment-related tags, corresponding to opposing poles of the three Osgood dimensions or Mehabrian states set out in section 2.1, a subset was selected of 5268 lexical items consisting mainly of an even distribution of modifiers, nouns and verbs which are reliably coded for 15 sentiment features listed in table 2, 2 activation, 2 dominance and 11 evaluation features . For clarity, this sub-lexicon will be referred to as GI_{sent} throughout this paper. The 15 features were deemed after hand

validation to consistently encode sentiment without redundancy and are of potential utility for sentiment analysis. It is worth noting that there is not

Table 2 GI sentiment tags

Dimension	Tag	Example	Tag	Example
Activation	Active	"abolish"	Passive	"accept"
Dominance	Strong	"admirer"	Weak	"afraid"
Evaluation	**Positive Valence**		**Negative Valence**	
	PosAff	"ardent"	Fall	"collapse"
	Positiv	"comedy"	Hostile	"combat"
	TrnGain	"afford"	NegAff	"condemn"
			Negativ	"conflict"
			Pain	"cramp"
			TrnLoss	"cut"
			Vice	"contempt"
			WlbLoss	"die"

an even distribution of lexical items encoded for the poles of emotion dimensions: more lexical items are encoded as negative, active and strong than as the corresponding positive, passive and weak categories. This distribution is consistent across part-of-speech categories with the exception of passive-active adjectives where there are more adjectives encoded as passive than active. The skew in frequency of polarity items is explored in greater depth in the corpus analysis in section 4.

3.1.2 Dictionary of Affect in Language

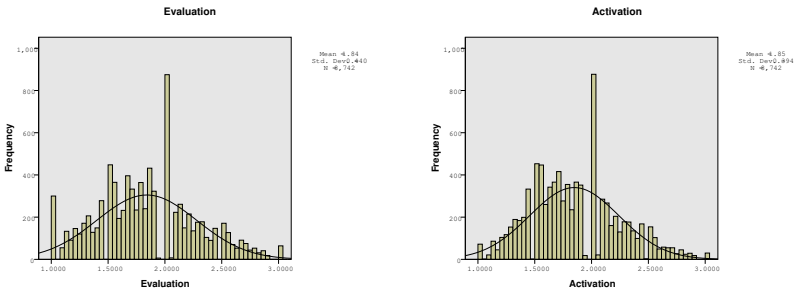
Provenance. The Dictionary of Affect in Language (DAL) is a hand-tagged, frequent word-list developed at the Laurentian University with the aim of providing a resource for the quantification of emotion in language (Whissell 1989). Like GI, it derives from the traditions of corpus-based lexicography and lexicalisation of emotion according to psychological principles. It is composed of a word list of the most frequent terms in the 1960s Brown Corpus cross-referenced with a 1990s corpus of texts by adolescent and young adult North Americans. The word list derived from the corpora was tagged by hand by approximately two hundred volunteers, with the final sentiment values assigned being an average of the 5 to 8 individual ratings per category for each word in the list (see Sweeney and Whissell (1984) for details of the validation process).

Representation of Lexical Items and Emotion. The DAL contains 8742 lexical items, including morphological variants of the same lemma (e.g. dog and dogs). Each item is tagged on a scale of 1–3 for the two Osgood emotion dimensions evaluation and activation and for an additional category, imageability, which we have not included in our analysis (see table 3 for examples). There is no relationship between the categories (i.e. they are orthogonal to each other ($\rho <$

Table 3 DAL examples

Word	Evaluation	Activation	Imageability
grief	1.2500	2.0000	2.0
joy	2.8889	2.3077	1.4
drawing	2.3077	2.3077	3.0

0.097)). This follows the representation of the two main emotion dimensions in the literature in figure 1. The Evaluation and Activation category values both follow a normal distribution, as in figures 2(a) and 2(b), both with mean value of circa 1.84, a little below the median value of 2. The Evaluation histogram shows an unusually large peak at the lower extreme of the evaluation scale $Val = 1$. This mirrors the predominance of negative tagging in the GI_{sent} lexicon and could suggest that there is a greater lexical variety for expressing negativity. Overall the two Osgood dimensions are represented by raters as quite polarised.



(a) Evaluation (b) Activation

Fig. 2 Histograms of DAL Categories

3.1.3 WordNet Affect

Provenance. The WordNet Affect (WNA) dictionary (Strappavara and Valitutti 2004) was developed as part of the WordNet Domains initiative at ITC-IRST, Italy from 2000 to annotate the WordNet knowledge base (Fellbaum 1998) with domain information according to the Dewey Decimal Classification system. WordNet itself is organised as a network of *word meanings* rather than *word forms*. The basic unit in the lexicon is the set of synonymous words, termed a Synset and a range of lexical and semantic relations, such as hyponymy and antomymy, can hold between pairs of Synsets. WN Affect is composed of those elements of WordNet which have been encoded with Affect domain labels following in the spirit of Ortony et al. (1987). Unlike the GI and

DAL, this resource is not corpus-driven but rather is derived from both intuitive and automated processes. An initial core set of 1903 emotion terms were annotated for the affective features listed in table 4. This wordlist was mapped onto corresponding Synsets and the list expanded to 2874 synsets in total. The expansion process was automated for WordNet relations which were deemed to preserve affect (e.g. similarity, derived-from, etc) and performed manually for other relations (e.g. hyponymy, entailment, etc).

Representation of Lexical Items and Emotion. WN Affect contains 2,874 WordNet synsets which encode 4,787 words (51% adjectives, 27% nouns, 11% adverbs and 11% verbs). It is fundamentally different from other lexica in that it does not represent emotion in terms of Osgood dimensions by assigning a polarity to the affective terms but rather encodes domains of emotional experience from inherent traits to more ephemeral responses, traits or situations. This lexicon provides an interesting counterfoil against which to evaluate other lexica as it represents a taxonomy of emotional experience rather than dimensions of emotionality in text. Fourteen affective domains are encoded within the lexicon. Of these, three are used for only 1–4 synsets (manner (man), words derived from emotion (psy) and state (sta)). The eleven widely used domains are set out in table 4 with examples. Approximately 3 in 10 terms have more than one affect domain assigned, up to a maximum of 6 affect domain assignments for the term "sick". The most commonly co-occurring domains are: attitude, behaviour, emotion and traits or cognitive and physical states and emotions. This resource represents a very different model of viewing emotion that may or may not be represented in the Osgood dimensional space. As such, their contribution to the affective contour of text and how an automated system should deal with them remains to be determined.

Table 4 WN Affect Domains

	Domain	Number	Examples
1	**Attitude (att)**	708	Intolerance, belligerent
2	**Behaviour (beh)**	484	approval, inhibited
3	**Cognitive State (cog)**	685	confusion, wistful
4	**Edonic signal (eds)**	105	carsick, gracious
5	**Emotion (emo)**	2045	anger, fear
6	**Mood (moo)**	71	animosity, amiable
7	**Physical State (phy)**	220	depression, alive
8	**Emotional Response (res)**	55	palpitation, livid
9	**Sensation (sen)**	126	pleasure, thirsty
10	**Emotion-eliciting situation (sit)**	282	quietude, vivacious
11	**Trait (tra)**	1598	superiority, itchy

3.1.4 SentiWordNet

Provenance. SentiWordNet (SWN) was developed at ISTI-CNR in Pisa, Italy since 2005 using the WordNet knowledge base (Fellbaum 1998) as its basis.

It is a very wide coverage resource which was generated automatically using a small hand-selected seed set of twelve unambiguously positive and negative terms (see below) to generate a sentiment rating for other terms in WordNet by propagating semantic links through the knowledge base. This analysis is based on SWN 1.0. SWN 3.0 has since been released. based on a more recent version of WordNet and using a revised training algorithm (Baccianella et al. 2010). The changes to the underlying WordNet version (2.0 to 3.0) are largely in the connectivity of the database (derivational morphology and domain links) rather than in terms of lexical coverage and so the findings of the analysis reported here in terms of lexical coverage still hold. As regards polarity annotation, version 3.0 of SentiWordNet is reported as up to 20% improved on version 1.0. This very valuable improvement does not represent a radical change in values but rather a fine-tuning of annotation. The overall thrust of the findings are applicable to both resources.

Representation of Lexical Items and Emotion. SWN is an overlay on WordNet and contains 28,428 WordNet Synsets (10263 adjective, 2455 adverb, 13150 noun and 2560 verb synsets) which include a total of 39,066 individual terms.[2] Each Synset is encoded with both a positive and negative sentiment polarity rating (*posSent* and *negSent* respectively) and an objectivity rating all between the values of 0 and 1 and summing to 1, as in the examples in table 5. The positive terms in the seedset set out in table 6 were assigned the maximum positive value $posSent = 1$ and minimum negative value of $negSent = 0$ and an objectivity value of 0. Likewise the negative seedset terms have $negSent = 1$ and $posSent = 0$. The distribution of positive and negative sentiment ratings

Table 5 SWN examples

Synonym list	**Positive rating**	**Negative Rating**	**Objective Rating**
casual, everyday	$posSent = 0.375$	$negSent = 0.125$	$obj = 0.5$
heartsick, heartbroken	$posSent = 0.0$	$negSent = 0.625$	$obj = 0.375$

Table 6 SWN seedset terms

$posSent = 1$	virtuous, upright, decent, fortunate, nice, good
$negSent = 1$	badness, denigrating, hapless, libellous, pathetic, negative.

(*posSent* and *negSent*) in the lexicon follow a Power Law distribution with $posSent = 0$ or $negSent = 0$ the most frequent rating (up to 30%), as shown in figure 3.1.4. The mean positive and negative ratings are $posSent = 0.1849$

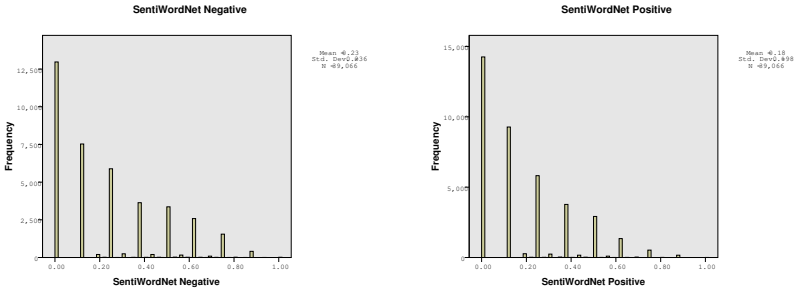
[2] SWN in fact is an encoding on top of the complete WordNet knowledge base of 115423 synsets but 86995 of these have no sentiment rating, i.e. $posSent = 0$ and $negSent = 0$. These non-sentiment synsets have been ignored for the purposes of this study.

and $negSent = 0.2326$ respectively, i.e. low sentiment ratings predominate. This tendency to higher negative ratings is consistent across most part–of–speech categories, although noun and verb mean ratings ($posSent = 0.14$, $negSent = 0.21$ and $posSent = 0.14$, $negSent = 0.18$ respectively) are slightly lower than adjectives ($posSent = 0.19$, $negSent = 0.27$). The mean ratings for adverbs however invert the trend with higher positive ratings for this 8% of the lexicon ($posSent = 0.35$ and $negSent = 0.1$). Strongly affective terms can be defined as those with polarity ratings greater than the median of 0.5, defined for the purposes of this analysis as $posSent$ *or* $negSent \geq 0.6$. In this subset of almost 24% of the lexicon, there is a predominance of negative terms, as in both the DAL and GI_{sent} lexica. Strongly affective negative terms account for 16% of the lexicon with a mean negative score of $negSent = 0.692$. Strongly affective positive terms account for only 7.4% of the lexicon with a mean rating of $posSent = 0.679$. Hence negativity is both more common (16% : 7.4%) and somewhat more pronounced (0.692 : 0.679) than positivity in the lexicon. Again, these findings are consistent across parts-of-speech with the exception of adverbs where strongly positive terms are 2.4 times more frequent than negative ones.

As the lexicon was automatically generated and was not validated by hand, it contains some errors. For example, the most positive terms include the terms "ill-mannered", "perverse", "sleazy" and among the top negative terms there are "gladsome" and "extralingusitic". These errors may be due to bugs or over-generation of rules in the classification process. Indirect antonymy relations for example seem to consistently lead to incorrect polarity assignments and the use of lexical negation with prefixes such as "un" and "non" for classification seems to over-generate. Furthermore, there is the problem of polysemy with WordNet encoding multiple fine-grained senses in the lexicon, including even ironic word senses at times. Despite these difficulties however, SWN constitutes a wide coverage lexicon with positive and negative polarity ratings for all terms, with a 20% improvement to ratings in SWN 3.0, and a very rich semantic basis provided by the WordNet conceptual hierarchy.

3.2 Lexical Content Overlap

As illustrated in section 3.1, the four lexica vary hugely in terms of their structure, encoding, conceptual underpinning and selection criteria. This section aims to examine how these very different approaches in fact impact on the contents of the lexica. The underlying research question here is whether the different lexica in fact represent a subset of language, a language of emotion, which is coherent and consistent regardless of the approach taken to the lexicon building task. The comparison of lexical content comprises two evaluations: an analysis of the contents and significance of the pair-wise overlap between lexica (section 3.2.1) and an analysis of the characteristics of the set of terms shared across all four lexica as a potential core of emotion-bearing terms (section 3.2.2).



(a) Negative (b) Positive

Fig. 3 Histograms of SWN Polarity Categories

3.2.1 Pair-wise Lexical Content Overlap

The overlap between lexica was determined based on shared orthography and part of speech where available. Unlike GI, SWN and WNA, the DAL lexicon does not encode part-of-speech and deals in full forms rather than lemmas. However, in most cases the base forms are available in the lexicon also. As the sentiment values for each full form of a lemma are not the same, we decided not to average over sentiment values but rather to look for exact matches of forms where possible. Table 7 sets out the number of terms shared between lexicon pairs, i.e the pair-wise intersections of the lexica. In order to estimate

Table 7 Overlaps between Sentiment Lexica

	GI_{sent}	**SWN**	**DAL**	**WNA**
GI_{sent}	5268	3851	2532	1397
SWN		39066	3868	4135
DAL			8742	1038
WNA				4603

the significance of the lexical intersections relative to the overall contents of the lexica, two similarity metrics were calculated, the Dice and Asymmetric coefficients. The similarity coefficients were calculated according to the formulae set out in equations 1 and 2, where $a = lexiconIntersection$, $b = size\ lexicon\ 1$ and $c = size\ lexicon\ 2$.

$$Dice = \frac{2a}{2a+b+c} \tag{1}$$

$$Asymmetric = \frac{a}{a+b} \text{ or } \frac{a}{a+c} \tag{2}$$

Both coefficients provide a measure of the significance of the intersection between the two lexica, however with different emphases. The Dice coefficient

represents the significance of the intersection of the two lexica relative to both lexica taken together. The Asymmetric coefficient on the other hand addresses the issue of possible differences in the cardinality of the two lexica, evaluating the significance of the intersection relative to a single lexicon only. Both metrics are reported as the Dice coefficient provides a general notion of overlap while the Asymmetric measure allows us to tease out the relationship between the lexica where the intersection may be significant relative to one smaller lexicon but insignificant relative to an other. For both coefficients, a value of 0 indicates no overlap and 1 complete overlap. For the purposes of this analysis values over 0.2 are considered of interest and will be discussed below. The coefficients for each lexicon pair are set out in table 8 with values of interest represented in bold.

A first point to note is the difference between the Dice and the Asymmetric coefficients, in particular in relation to SWN intersections. In fact, the Dice coefficient gives a somewhat distorted view of the degree of overlap between SWN and the other lexica. According to the Dice coefficient, the intersection of SWN with all other lexica is low ($0.161 < dice < 0.189$), similar to the Asymmetric coefficient values relative to the SWN lexicon (table 8, row 6): $0.099 < asym < 0.106$. However, SWN is by far the largest lexicon, containing 39,066 lexical items and the coefficient values reflect this asymmetry of size with respect to the other lexica, evaluating the intersection relative to the hugely dominant SWN rather than to both lexica. However, the Asymmetric coefficients relative to the other lexica (table 8, column 2) give a clearer picture. The results here are altogether different showing a major contribution of the lexicon intersection relative to the other lexica in the lexicon pairs ($0.442 < asym < 0.898$). These results suggest that the SWN lexicon subsumes from 44% to 89% of the other three sentiment lexica, in terms of its lexical content at least. These findings raise some questions about the composition of SentiWN: while it does include many lexical items derived from psycholinguistic experimentation, it also includes many hundreds and thousands of lexical items which have not been suggested in previous interrogation of human subjects or corpora. The lexicon therefore has the widest coverage but the reliability of *all* its lexical items may be questionable.

Table 8 Similarity Coefficients for Sentiment Lexica Pair-wise Intersection

		1	**2**	**3**	**4**
	Dice	GI_{sent}	**SWN**	**DAL**	**WNA**
1	GI_{sent}	1	0.174	**0.361**	**0.283**
2	SWN		1	0.162	0.189
3	DAL			1	0.155
4	WNA				1
	Asymmetric	GI_{sent}	**SWN**	**DAL**	**WNA**
5	GI_{sent}	1	**0.731**	**0.481**	**0.265**
6	SWN	0.099	1	0.099	0.106
7	DAL	**0.290**	**0.442**	1	0.119
8	WNA	**0.303**	**0.898**	**0.226**	1

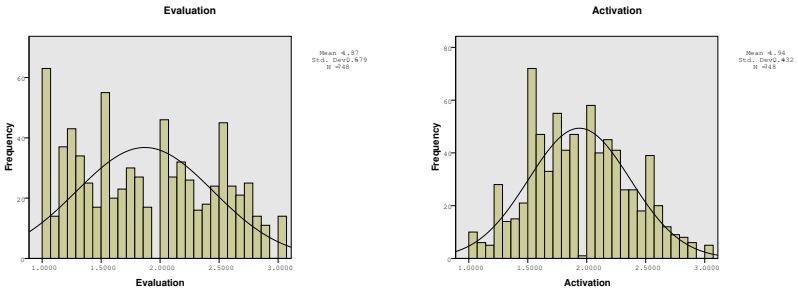
As regards the remaining three lexicon, the degree of overlap varies considerably and again the comparison of the Dice and Asymmetric measures is in some cases enlightening. The overlap between the GI_{sent} and WNAffect lexica is substantial ($dice = 0.283$). The two relevant Asymmetric coefficients confirm that this intersection is of medium importance to both lexica with the intersection accounting for 26% and 30% of GI_{sent} and WNAffect respectively. The Dice coefficient would also suggest substantial overlap between the GI_{sent} and DAL lexica ($dice = 0.361$). The Asymmetric coefficients allows us to tease out the apparently strong relationship between the GI_{sent} and DAL lexica. With respect to the GI_{sent} lexicon, the intersection is very important accounting for almost half of GI_{sent} ($asym = 0.481$). However, the overlap only accounts for 29% of the DAL, an important but nevertheless weaker contribution. It is interesting to note that the asymmetric coefficient of the full GI_{sent} lexicon (sentiment and non-sentiment bearing terms) with respect to DAL is in fact much greater, with the number of intersecting terms accounting for 50% of the DAL. Given that both lexica are based on American English frequency lists, psycholinguistic experimentation and introspection of human respondents, it is interesting that although the intersection of the two lexica is high for the full GI lexicon, it is not just the sentiment features of that lexicon that are responsible for this intersection (illustrated by the large drop in intersection from full GI to GI_{sent}, $fullGI \cap DAL = 4424$ to $GI_{sent} \cap DAL = 2532$). The divergence could be due to the different time periods on which the word lists for the two lexica are based or the impact of the corpus filtering of the DAL word list. In the absence of a diachronic study of sentiment in language this hypothesis cannot be confirmed. Finally, the Asymmetric coefficient results highlight a weak relationship between the DAL and WNAffect dictionaries, where 22% of WNAffect overlaps with the DAL. This relationship is insignificant relative to the DAL, as reflected in the Asymmetric score ($Asymm = 0.119$).

In summary, in a pair-wise comparison of the lexical items in the four sentiment lexica, SWN has the widest coverage and subsumes between 50% and 90% of the other three lexica. The contributions of other lexica however do not account for a significant portion of SWN itself, therefore the accuracy of this automatically generated lexicon may be somewhat in doubt. The intersection of the other three lexica is not negligible, nor is it very significant. The GI_{sent}, DAL and WNAffect lexica have some shared information content but each merit examination and use on their own as the shared information is only at most 50% of a given resource. This is especially true in the case of the DAL which seems to diverge most in terms of lexical content from the other three with only 11-44% of the lexicon subsumed in other resources. As a high frequency word list this might be expected although the GI lexicon also maintains this characteristic.

3.2.2 Shared Content in All Lexica

In total, there are 748 lexical items which are shared by all four lexica. While this is not a significant proportion of any lexicon, it is enlightening to examine

the characteristics of the lexical items common to all lexica to determine if these are the sentiment-bearing core of the lexica, those quintessential terms which unequivocally encode sentiment. This section sets out a comparative analysis of the sentiment feature distributions of these 748 terms relative to the full lexica. As regards the DAL, a comparison of the histograms in figure 3.2.2 suggest that the 748 shared terms are not representative of the lexicon as a whole. The values for all scales, in particular the evaluation scale, are spread more evenly across the spectrum, with a lower peak at the median value of 2 and more terms in the tails of the distribution (lower and higher values). The 748 terms represent a more evenly distributed sample of the lexicon across the evaluation and activation spectrum of values. The SWN values also are no

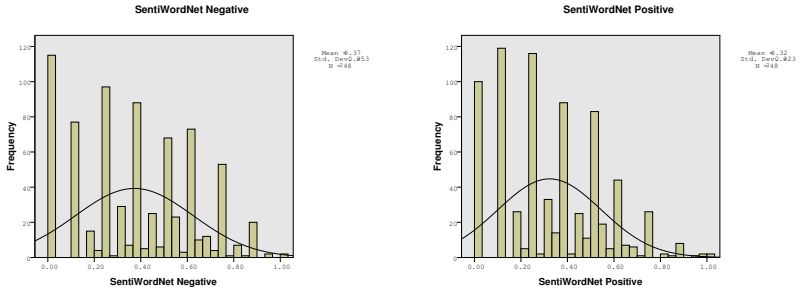


(a) Evaluation (b) Activation

Fig. 4 Histograms of DAL Categories for 748 Overlap Terms

longer normally distributed but skewed towards the lower end of both the positive and the negative scales, as in figure 3.2.2. This result would suggest that the 748 shared term sample has a higher proportion of less strongly positive and negative terms than the lexicon as a whole. As regards the GI_{sent} lexicon, the proportion of positive to negative polarity features is significantly different between the overlap and the lexicon itself ($\chi^2(1, N = 8641) = 69.4, p < 0.005$), with the ratio of negative to positive smaller than for the full lexicon (1:2, rather than 1:3). This would suggest that the shared features are somewhat more evenly distributed between positive and negative terms. In the case of the WN Affect dictionary, the distributions of WNA domain tags are not significantly different from that of the full lexicon, with the same proportion of emotion experience types represented.

In summary, for the lexica that encode polarity, the overlapping terms tend toward a more even distribution of polarity values than in the full lexica. They do not seem to encode the extremes of the sentiment poles but neither do they encode a generic middle ground. Rather they represent a selection of terms common to all resources which cover the full spectrum of sentiment values,



(a) Negative (b) Positive

Fig. 5 Histograms of SWN Polarity Categories for 748 Overlap Terms

from low to high, reflecting perhaps the frequent word list basis of at least the GI and DAL lexica.

3.3 Sentiment Assignment Evaluation

The degree of overlap and consistency in terms of lexical coverage was outlined in section 3.2. The focus now shifts to correlations between the sentiment values assigned to these lexical items shared between lexica to evaluate the consistency of the sentiment features across sentiment resources. Section 3.3.1 outlines the feature transformations required to derive comparable sentiment features across the four lexica which differ in their respective representations of sentiment values. Section 3.3.2 outlines the key findings of the pair-wise comparisons of GI_{sent}, SWN and the DAL and the implications of these for resource selection in Sentiment Analysis. As noted in section 3.1.3, the WN Affect lexicon differs fundamentally from the other lexica in terms of the features it encodes, an analysis of the feature assignments of this lexicon relative to the other three is therefore dealt with separately in section 3.3.3.

3.3.1 Comparability of Sentiment Features

As set out above, GI_{sent} assigns binary categorical tags to sentiment-bearing terms while the DAL and SWN assign scale values indicating sentiment intensity and orientation for all terms in the lexica. SWN and DAL may be compared directly using a correlation metric as they both use scale values for feature assignment. An analysis of the level of agreement in sentiment assignments with respect to GI_{sent}, however, requires that scale values are represented as binary tags or *vice versa*. For the purposes of this analysis, therefore, the scale values of DAL and SWN were converted to binary categorical tags for comparison with the GI_{sent} negative and positive categories in χ_2 tests for independence.

Two sets of binary variables were generated from two different transformations for the DAL and SWN scale ratings in order to ensure that the results were not an artefact of the transformation type. As regards SWN, the first transformation takes sentiment polarity relative to the lexicon as a whole, sentiment value relative to the mean sentiment assignment in the lexicon ($binaryMean$). The second takes sentiment polarity as a term-internal value relative to the opposite polarity sentiment assignment for that term ($binaryLarger$). As regards the DAL, sentiment polarity was based on the evaluation assignment for a term relative to either the evaluation mean, representing lexical distributions of sentiment values or relative to the evaluation mid-point (2), representing an absolute sentiment neutral point on the evaluation scale. The transformation formulae for both SWN and DAL are set out in table 9.

Table 9 Binary Variables derived from SWN and DAL scale values

Lexicon	Variable Name	formula
SWN	**binaryMean_Sent**	$SWN\ Sent > SWN\ Sent\ Mean$
	binaryLarger_Sent	$SWN\ Sent > SWN\ OppositeSent$
DAL	**binaryMean**	$DALVar > DALVarMean$
	binaryMidPoint	$DALVar > 2\ (DAL\ MidPoint)$

3.3.2 Sentiment Assignment Evaluation: Key Findings

Statistically Significant Polarity Agreement. In a pair-wise comparison of polarity values in the three lexica, GI_{sent}, SWN and DAL, we found for all lexica there is statistically significant agreement in the polarity assignments of the overlapping terms. For the comparison of GI_{sent} with SWN and DAL, binary polarity tags in the lexica were compared in a χ_2 test for independence. For all variable pairs, the null hypothesis that there is no relationship between the binary polarity assignments for each lexicon can be rejected at $p < 0.0001$, as illustrated in table 10. This relationship is positive for same polarity pairs and negative for opposite polarity pairs.

Agreement in polarity assignments between SWN and DAL were compared by computing the Pearson Correlation coefficients for the scale sentiment variables of shared lexical items. As might be expected, the Evaluation dimension is correlated with the SWN values sentiment values, negatively correlated with SWN_neg ($r = 0.212$, $p \leq 0.01$) and positively correlated with SWN_pos ($r = 0.264$, $p \leq 0.01$) . These correlations are statistically significant at $p \leq 0.01$ but not very strong, suggesting that though the relationship is present, it is weak. Therefore, as DAL evaluation values decrease, SWN negative values increase marginally and as DAL values increase, SWN positive values increase marginally.

Table 10 χ^2 coefficients for DAL Evaluation and GI_{sent} Polarity Features, significant for $df = 1, p < 0.0001$

DAL Evaluation	GI_{sent} **Negative Features**	GI_{sent} **Positive Features**
binaryMean	$\chi_2 = 726.004$	$\chi_2 = 709.605$
binaryMidPoint	$\chi_2 = 624.806$	$\chi_2 = 601.206$
SWN	GI_{sent} **Negative Features**	GI_{sent} **Positive Features**
binaryLarger_Neg	$\chi_2 = 562.439$	$\chi_2 = 452.742$
binaryLarger_Pos	$\chi_2 = 509.970$	$\chi_2 = 471.513$
binaryMean_Neg	$\chi_2 = 378.114$	$\chi_2 = 177.324$
binaryMean_Pos	$\chi_2 = 219.495$	$\chi_2 = 404.028$

Other Sentiment Dimensions: little correlation. The polarity of the evaluation sentiment dimension seems to be consistent across lexica. A secondary analysis aimed to investigate the consistency of how other sentiment dimensions (activation and dominance) are represented. In the two lexica which explicitly encode the activation dimension (GI_{sent} and DAL), there is some relationship between the GI_{sent} variables and all three DAL dimensions. However, in a comparison of the distributions of the GI_{sent} activation and dominance features (active, passive, strong, weak), there was only a very weak relationship with DAL variables, including the DAL activation feature. A further analysis was carried out to investigate the hypothesis that the intensity of SWN polarity scaled values conflate evaluation and activation dimensions and therefore can be approximated by examining a combination of evaluation and activation features. A principal components analysis of the SWN polarity features and the GI_{sent} evaluation, activation and dominance features was carried out. The principal components detected only accounted for 37% and 43% of the variance of the SWN features. Similarly, the Pearson correlation coefficients computed for SWN polarity and DAL Activation features strongly suggest that SWN polarity values are completely uncorrelated with the DAL Activation dimension (for SWN negative: $r = -0.023$; for SWN positive: $r = 0.007$; not significant at $p \leq 0.01$). These results suggest that the scaled positive and negative values in SWN are not in fact a conflation of Osgood's evaluation and activation, at least as they are represented in GI_{sent} and DAL but rather that there is a scale of positivity and negativity for lexical items which is unrelated to other emotion factors. The correlation coefficient is very low and for this reason, despite the potential for error in SWN, this result is not likely to be due to chance. Furthermore, we can conclude that, where the lexica overlap, although the positive and negative evaluation assignments are quite consistent, other Osgood dimensions are either not represented or where they are, they are not represented consistently.

3.3.3 WN Affect correlations

As noted in section 3.1.3, the WN Affect lexicon differs fundamentally from the other lexica as, although individual lexical items are tagged for membership of affective domains, they are not assigned an explicit polarity. For this reason,

it is not possible to examine direct correlations between sentiment polarity in WNA and other lexica. However, as the intersection between WNA lexical items and other lexica is quite significant with respect to WNA (see table 8), it is possible to investigate the dominant polarity of the different WNA domains and possible correlations with other lexicon categories. This section teases out some of the latent characteristics of the WNA domains in terms of their polarity, activation and levels of abstraction.

An analysis of sentiment feature correlation between WNA and other lexica reveals very interesting latent polarity characteristics of WNA domains. Firstly, WNA domains appear to each have a dominant polarity which corresponds to a distinction between long-term aspects of emotional experience (traits and attitudes) and short-term ones (responses and behaviour). This distinction is statistically significant for the three lexicon overlaps. In χ^2 tests of independence exploring polarity assignments for terms shared with GI_{sent}, the distribution of positive and negative in certain WNA domains was significantly different from the overall lexicon to warrant mention. Both the attitude and trait domains contain significantly more positive than negative terms at $p < 0.0001$, while the cognitive state, emotion, mood, physical response and response domains are significantly more negative than positive. This division of WNA domains roughly corresponds to a notion of long–term tendencies *vs* short–term responses. Similarily in an analysis of terms shared with SWN, for the predominantly "short–term" domains (edonic signals, emotions, mood, physical response, response, sensation, situation), the tendency was replicated with significantly more strongly negative and significantly less strongly positive lexical items. In the case of the behaviour and manner domains, only a tendency towards less positivity was noted. However, for the long–term attitude and trait domains, the tendency is reversed with more positive and less negative terms in both. The same distinction between WNA long-term trait and short-term response domains is replicated in an analysis of DAL shared terms where the long-term domains (attitude, trait) show a statistically significant trend towards less negative and more positive terms and the converse for some short-term domains (cognitive state, emotion, response). The polarity findings suggest that although long-term tendencies may be predominantly positive, the more short-term responses tend to be negative or there are many more negative ways to describe or enumerate them.

Secondly, not only is there a polarity bias but there is some evidence for an activation and imageability bias in some WordNet Affect domains which correspond to intuitive categorisation of emotional experience as physical or cognitive and internal or external. For the activation dimension there seems to be a polarisation of values where shared lexical items are either more strongly active or strongly passive, approximately 10% more in both cases. The WNA domains attitude, cognitive state, mood, situation and trait show a statistically significant tendency towards more passive terms while the domains behaviour and emotion tend towards more active terms. This could reflect a distinction between physical activities or responses and more passive cognitive, latent concepts or features. As regards the imageability dimension, there would seem

to be a trend towards more concrete terms in the WNA lexicon with 10% more concrete terms and 10% less abstract terms than usual in the full DAL. The cognitive state and trait domains have significantly more abstract terms while the emotion and response ones have significantly more concrete terms. This again reflects a division between internal representations and external manifestations of emotion which could be conceptualised in terms of abstract and concrete.

3.4 Conclusions and Implications for Sentiment Analysis

The detailed analyses of sentiment lexica set out above illustrate that the lexical coverage of the lexica is quite varied. Although they do overlap in some of the terms covered, there is no coherent pattern to the overlap regardless of the theoretical underpinning or mode of development of the lexica. As regards the representation of sentiment within the lexica, where there is lexical overlap, they are consistent in sentiment polarity assignments. However, there is little consistency or relationship between other features which the various lexica encode. Finally, the WordNet Affect lexicon shows distinct polarity biases for different domains of emotional experience. The impact of these biases or of a possible underlying distinction in how emotion is experienced over time could have an impact on sentiment analysis applications which remains to be explored.

4 Corpus Analysis of Affective Language

The previous section examines the lexical resources of emotion in terms of their consistency and coverage relative to each other. This section provides an analysis of their use and distribution patterns in the English language in general, represented by the British National Corpus (BNC), and in sub- or special languages of English, represented by the BNC Imaginative and Informative sub-corpora and a separate corpus of financial news text. This analysis addresses the issue of whether the lexica of sentiment constitute a coherent subset of the English language with usage patterns that set them apart from general language. Although no lexicon can be fully comprehensive, this analysis would lend support to the lexica as repositories of the language of emotion. Secondly, the aim is to determine the usefulness of each lexical resource in terms of the distribution of their lexical items and features, in particular polarity features, in general language. While the realisation and interpretation of sentiment in text is a very complex phenomenon where individual sentiment-bearing lexical items are only one factor in the complex interplay of textual elements (Martin and White 2005; Polanyi and Zaenen 2004), this analysis provides an insight into how prominent is the lexical basis for sentiment as represented by the sentiment lexica in text in general language. It is invaluable to those working in automatic sentiment analysis to select between available

resources and potentially specify requirements for new or extended lexical resources. The further comparative corpus analysis aims to determine whether lexical and polarity distributions differ across registers or sub-languages of English. This more focused analysis highlights the domain–dependent nature of affect in text and the possible need to re–assess resource requirements, in particular the need for domain–specific resources, for different sentiment analysis applications. Section 4.1 sets out the details of the lexical features under investigation in this corpus analysis and some issues with inter-lexicon consistency. The corpus analysis of affective text in general language is outlined in section 4.2 and the comparative corpus analyses in section 4.3.

The findings would strongly suggest that the sentiment lexica examined in this article do constitute a statistically distinct subset of English. Furthermore, our analyses would suggest that there may be a positivity bias inherent in language which needs to be accounted for in SA systems

4.1 Lexical Features and Frequencies for Corpus Analysis

The corpus analysis presented here is based on the four lexica presented in the previous section. For each of the four sentiment lexica under investigation, the frequency count of each lexical entry was calculated in the full BNC, the imaginative and informative BNC sub-corpora and the financial corpus. These term frequencies were used to determine sentiment term distributions and relative coverage of each lexicon for the different corpora. In addition, the distributions for the sentiment features encoded in the four lexica (listed table 12) were calculated based on these term frequencies. The feature counts and distributions are evaluated to determine which features are most salient or dominant and to compare feature occurrence across corpora. It should be noted however that there are a number of issues related to the derivation of the basic term frequency counts which require some comment. Firstly, there is the question of orthographical consistency between the four sentiment lexica which are of American origin and the BNC which is predominantly a repository of British English. Of the 4 sentiment lexica under investigation, the two based on WordNet (SWN and WN Affect) include both UK and US orthography and can therefore be used in this corpus analysis without modification. The GI_{sent} and DAL lexica, on the other hand, had to be modified to include UK orthography in order to carry out the corpus analysis. Secondly, the lexica are not consistent in their representation of lexical items: DAL uses full lexical forms while the other three use lemmas. In essence, the sentiment features provided in the DAL lexica claim only to hold for individual lexical forms, not for all forms of lemmas, while the other three lexica make the assumption that features hold across all forms of a lemma. The use of the WFWSE BNC frequency lists (Leech et al. 2001), in fact, solves this potential problem as it provides both lemma and full form counts for all BNC lexical items and so the assumptions of both lexicon types can be upheld, full form counts are used for DAL and lemma counts for the other lexica. Thirdly, a further

discrepancy between the lexica is their provision of part–of–speech (POS) tags for lexical items. Again, the DAL is distinct in that it does not provide POS tags for lexical items. Therefore, although counts are based on full forms, the full forms are not disambiguated for part–of–speech. As the corpus frequency lists do include POS tags, the DAL frequencies are the sum of frequencies for all parts-of-speech for any given full form. The other lexica counts are based on lemma counts for the lexical part–of–speech only. The corpus frequency lists do not include any multi-word lexical items whereas the lexica do include some multi-word entries. Multi-word lexical entries were in effect ignored for the purposes of this analysis as the number of multi-word entries in all of the lexica is not substantial. Finally, the GI, WNA and SWN resources encode multiple word senses for some terms. As the BNC is not disambiguated for word sense, the multiple word senses are amalgamated. This of course entails a substantial loss of information carried in the lexica. This is issue is not resolved here but raised as an ongoing issue for SA systems using these lexica but not leveraging the additional disambiguated word information therein.

As regards the sentiment features for analysis, as noted in section 3.3.1, for some analyses it was necessary to transform the scale sentiment features in SWN and the DAL to binary variables such as those in GI_{sent} for comparability. The binary values were calculated according to the equations in table 11. In analyses that take account of the intensity of polarity values, the scale values of SWN and DAL are used directly.

Table 11 Binary Variables derived from SWN and DAL scale values

Lexicon	Variable Name	formula
SWN	**binaryLargerSent**	$SWN\ Sent > SWN\ OppositeSent$
DAL	**binaryMean**	$DALVar > DALVarMean$

4.2 Affective Text in General Language

4.2.1 General Language Corpus

The British National Corpus was used as the general language corpus in this study. This decision was motivated by the size (100 million words), broadness of coverage (10% spoken and 90% written text across a range of topics and registers) and accessibility of the corpus. In fact, the analysis is based on the BNC term frequencies as published in Leech et al. (2001) and available on-line at `http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html`. The term frequencies are provided as alphabetical lists of both lemmas and full forms with part–of–speech tags and frequencies reported per million words of the BNC.

Table 12 Lexical Sentiment Features

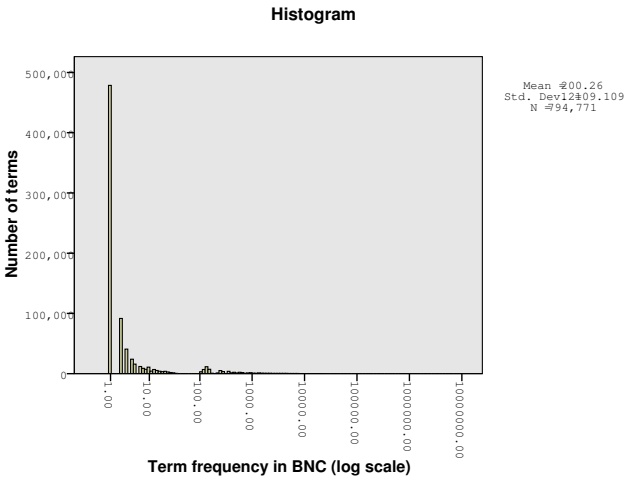
Lexicon	**Feature**	**Lexicon**	**Feature**
GI_{sent}	Fail	**WN Affect**	Attitudes (att)
	Fall		Behaviour (beh)
	Hostile		Cognitive State (cog)
	NegAff		Edonic signal (eds)
	Negativ		Emotion (emo)
	Pain		Mood (moo)
	PosAff		Physical State
	Positiv		Emotional Response (res)
	TrnGain		Sensation (sen)
	TrnLoss		Emotion-eliciting situation (sit)
	Vice		Trait (tra)
	WlbLoss		
SWN	swn_neg	**SWN**	swn_pos
DAL	Activation neg	**DAL**	Activation pos
	Evaluation neg		Evaluation pos

4.2.2 Sentiment Lexicon Term Frequency Distributions

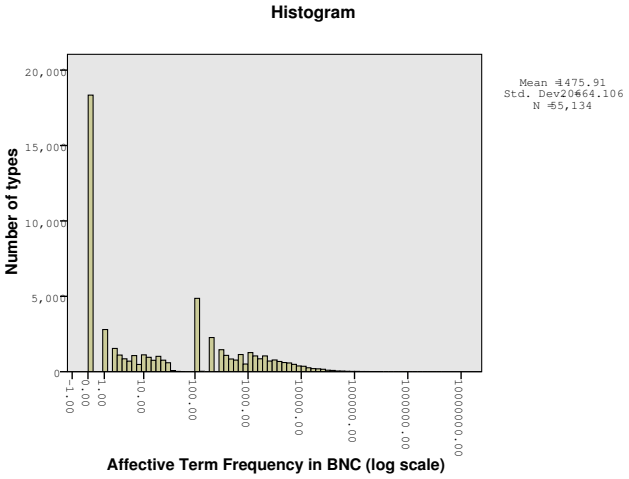
An initial analysis aims to identify if sentiment-bearing terms, i.e. those contained in sentiment lexica, have a unique distribution relative to general language, in terms of their frequency of occurrence in the BNC. The null hypothesis in question is that sentiment–bearing lexical items are no more or less frequent than other terms of the English language and do not have a distinctive distribution in language. The alternative hypothesis is that sentiment-bearing terms behave differently from general language and constitute a separate and specialised vocabulary of English. To test these competing hypotheses, the affective term distributions were compared with that of general language. The findings strongly suggest that the sentiment lexica do constitute a coherent and distinct subset of the English language.

Distribution Type. The distribution of term frequencies in the BNC, as in any large sample of natural language, is a Zipfian or Power Law distribution, as shown in the histogram in figure 6(a). According to Zipf's law, term frequency is inversely proportional to its rank in a frequency table, i.e. a few terms occur very often while the vast majority of terms occur very rarely. The frequency distribution for the combined sentiment lexicon terms follows the same distribution (figure 6(b)), as do the sentiment term frequencies for the individual lexica. The shape of the distributions is, therefore, the same but what of its size and spread?

Comparison of Means: Student's t-test. In order to estimate the similarity of the two distributions, we looked at a measure of central tendency, the mean term frequency. Table 13 sets out the mean term frequency and standard deviation for the full BNC, the combined lexica and each of the individual lexica. The table illustrates that the average frequency of terms is very different between the BNC and the sentiment lexica and between the lexica themselves.



(a) Full Term Frequency Distribution



(b) Affective Term Frequency Distribution

Fig. 6 Zipfian Term Frequency Distributions from BNC (log scale)

The Student's t-test provides a standard test to determine whether this difference in sample means is statistically significant. The null hypothesis here is that the means of the populations from which the two samples were taken are equal. In all cases, the null hypothesis could be rejected at $p < 0.0001$. This result should support the hypothesis that the sentiment lexica, both in com-

Table 13 General and Affective Type Frequency Mean and Standard Deviation in BNC

Corpus/Lexicon $N = 100,000,000$ **(BNC)**	**No of Types** (b)	**Mean Freq** $\mu = \frac{(b)}{N}$	**StdDev** $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
BNC Types	794771	150	11000
Combined Lexica	55134	1476	20664
GI_{sent}	10394	3752	45798
DAL	8671	4421	13131
WNAffect	4785	1493	20973
SentiWN	40619	720	22768

bination and alone, constitute a distinct subset of English with distribution parameters which are statistically significantly different from general language. More precisely, the mean frequency of sentiment terms is substantially higher than general language terms represented by the full BNC than would be expected according to chance. The sentiment-bearing types are between 5 and 30 times more frequent than other general language types suggesting that the language of "emotion" is very prominent, counting among its constituents some of the most common terms in English.

However, this term frequency data does not satisfy all the assumptions of the t-test and therefore its results may not be reliable. The sample sets are not normally distributed, as noted above, nor do they have equal variance. The test may be robust to a departure from these assumptions if the sample size, N is large enough as the standard error of the mean decreases and indeed here, N is very large ($N = 794771$ and $N = 55135$). However, given the violation of both the equal variance and normality assumptions, a further non-parametric test was performed to validate results.

Bootstrap Sampling Distribution To ensure that the mean term frequency of the sentiment lexica is not in fact representative of the BNC and that sentiment-bearing terms constitute a distinct and statistically different and highly frequent subset of English, the mean frequency was compared to a bootstrap sampling distribution of term frequency mean from the full BNC (Efron 1979). The sampling distribution was generated by taking 1000 random samples of terms (with replacement) from the BNC with a sampling size of $N = SentLexiconSize$ and the mean term frequency for each of the 1000 random samples was calculated. For a confidence level of $p < 0.05$, the null hypothesis is that the observed sample falls within 95% of the bootstrap random sampling distribution of means, not in the tails below the 2.5^{th} or above the 97.5^{th} percentile. More precisely, the null hypothesis is that the term frequency mean of the sentiment lexicon is representative of the population from which the bootstrapped distribution was sampled, i.e. the BNC or general language. The sampling distribution minimum and maximum sampled term frequency mean values and 2.5 and 97.5 percentile values for the different values of N are set out in table 14 along with the observed means for comparison. In all cases, the mean term frequency in the lexica is well outside 95% distribution

of randomised sampled means.[3] We may therefore reject the hypothesis that the sentiment lexica are representative of the full BNC population at $p < 0.05$.

Table 14 Bootstrapped Sampling Distribution of Mean Term Freq in BNC

Lexicon	Sample Size N	Lex Av Freq	Min	Max	2.5%	97.5%
All Lexica	**55135**	**1476**	70	498	78.34	318.78
GI_{sent}	**10394**	**3752**	44	1048	60.36	534.45
DAL	**8671**	**4421**	43	1094	58.64	529.69
SentiWN	**40619**	**720**	70	460	80.86	301.46
WNAffect	**4785**	**1493**	41.1	1511	49.96	718.1

Implications for Lexicon Selection. The tests set out above establish that the sentiment lexica constitute a distinct and very common subset of general language English as represented by the BNC. What then is the differential contribution of the individual sentiment lexica and the implications of this analysis for lexicon selection? The mean term frequency parameters for each lexicon combined with the lexicon size provide an indication of the characteristics of the different resources. The SWN lexicon, for example, is a wide coverage resource with over 40,000 terms. However, the average term frequency in SWN is only 720, greater than the full BNC average frequency but much lower than all the other sentiment lexica. This would suggest that while the term coverage is useful in theory, in practice, many of the terms may be encountered in free text only rarely. This wide but sparse coverage combined with the fact that it is an automatically generated dictionary which was not fully validated by human annotators would suggest that this resource may not provide as comprehensive coverage as its size suggests and it could be advisable to use it in combination with others for automatic sentiment identification in free text. In contrast, the DAL and GI_{sent} lexica which are each approximately a quarter of the size of SWN have a much higher mean term frequency. Furthermore, they are hand-built lexica, designed on sound psychological experimentation principles. These resources, although smaller, could prove as or more valuable than their larger counterpart. As noted in section 3.1.3, the content of the WN Affect lexicon is fundamentally different from the other three lexica in that it encodes aspects of emotional experience rather than emotional intensity or polarity ratings. For this reason alone, it is a valuable resource in itself. In addition, although the smallest of the lexica under investigation, the average term frequency for WN Affect terms is high (almost 1,500) and therefore its coverage is extensive despite not being broad.

[3] Only the WNAffect mean term frequency falls below the maximum sampled mean however this remains well outside the 97.5^{th} percentile.

4.2.3 Sentiment Polarity Feature Distributions

Given that the sentiment lexica constitute a distinct, non-random, highly-frequent selection of lexical items within general language, as suggested by the tests in section 4.2.2, this section aims to determine how the sentiment features in these lexica are distributed in natural language. Figure 7 illustrates the positive and negative feature counts per million words in the BNC for the three lexica which encode polarity (GI_{sent}, DAL and SWN), not normalised for lexicon size. Results strongly suggest a uniform tendency towards positivity in the BNC, regardless of lexicon. This finding supports the Pollyanna Hypothesis put forward by Boucher and Osgood (1969) where they showed that across languages and ages words at the positive end of the evaluation dimension were more frequently used than negative ones, even though there may exist more terms to express negativity, as this paper suggests is the case for the lexica in question. The choice of lexicon affects the degree to which this polarity bias is realised, with GI_{sent} positive and negative features giving the most extreme polarisation (positive:negative, 1:0.64) and SWN the least (1: 0.95). Although the bias may not appear very pronounced for some lexica, in all cases, the difference in proportions of positive to negative polarity values is significant (i.e. greater than would be expected according to chance according to a χ^2 tests of independence). The prevalence of positive terminology in general language may be associated with a general positive tendency, identified as a basic and universal characteristic of human nature and the positive:negative ratio of a lexical resource may have major implications for a sentiment analysis application in terms of how accurately it represents this basic characteristic.

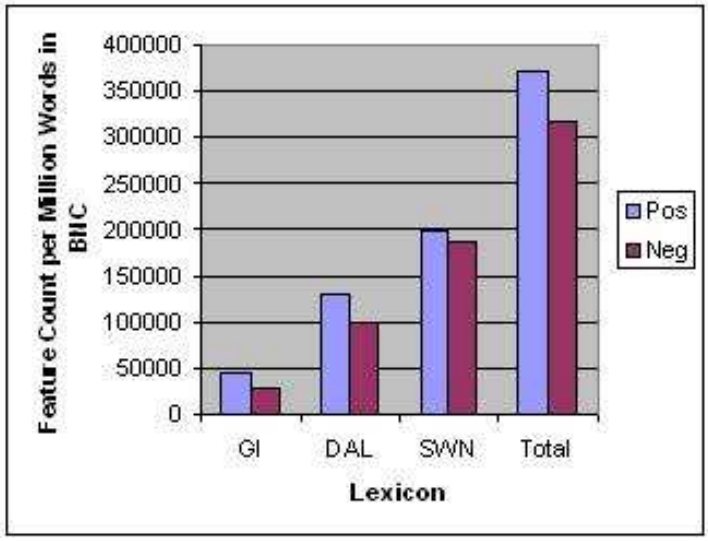


Fig. 7 Ratio of Positive to Negative Features in BNC for GI_{sent}, DAL and SWN

Both DAL and SWN give not only binary polarity tags but intensity ratings for each term. To test the hypothesis that polarity intensity values could invert the polarity bias in general language, for example, negative ratings are less frequent but more intense and therefore negativity could predominate, the polarity ratios for SWN and DAL were recalculated taking account of polarity intensity. Table 15 illustrates that intensity of ratings does impact on the positive and negative values and hence ratio with a relative increase in negativity and this difference is in fact statistically significant ($\chi^2 = 158.51$ and $\chi^2 = 63.64$ for SWN and DAL proportions respectively at $df = 1, p < 0.0001$). However, the difference is not large enough to invert the basic positive:negative ratio whereby positivity is predominant in the BNC with negativity strongly present but always with lower frequency. Given that the intensity of polarity items is difficult to assess out of context as evidenced by relatively low inter-rater agreement on this measure, the polarity intensity inversion hypothesis could be better tested by looking at the extent and persistence of impact of individual negative polarity items in text with human annotators, as suggested in (Devitt and Ahmad 2007), rather than relying solely on lexical distributions in corpora.

Table 15 Ratio of Positive to Negative Features incorporating Rating Intensity

Lexicon	Positive	Negative	Pos : Neg
DAL (incl. intensity)	41286.33	33940.05	**1 : 0.82**
SWN (incl. intensity)	72314.09	73876.47	**1 : 0.97**

Implications for Sentiment Analysis Applications. The polarity bias shown in the general language corpus is the opposite of the lexical polarity bias in each dictionary, set out in the Lexicon count column of table 16, where negative lexical items outnumber the positive. Although there appear to be more means of enumerating negativity in English as represented by the sentiment lexica, terms with negative polarity are used more sparsely in general language than positive. Table 16, showing the mean type frequency for each lexicon polarity category, bears out this finding. Although there are more negative than positive types in the sentiment lexica, in the corpus of general language they are on average 1.4 to 1.9 times less frequent than their positive counterparts. Whether this corresponds to negative sentiment being expressed more sparsely than positive sentiment in English or whether greater usage of positive lexical items is actually perceived as conveying greater positive sentiment are unanswered questions. A large-scale analysis of human responses to text is required in order to establish a polarity baseline for English or varieties of English and how sentiment is interpreted relative to this baseline, as noted in (Devitt and Ahmad 2007).

Table 16 Mean Type Frequency and Polarity Ratios for Lexical Polarity Categories

Lexicon	Polarity	Lexicon Count	Mean Freq in Corpus
GI_{sent}	**Positive**	1664	2432
	Negative	2010	1285
	Ratio	**1:1.21**	**1:0.53**
DAL	**Positive**	2523	5011
	Negative	5344	3437
	Ratio	**1:2.12**	**1:0.69**
SentiWN	**Positive**	16989	778
	Negative	19639	535
	Ratio	**1:1.16**	**1:0.69**

4.3 Comparative Corpus Analysis

Having explored how sentiment–bearing terms are distributed in general language, this section sets out to examine whether affective language use is homogeneous across different varieties and domains of language. Section 4.3.1 compares two varieties of English writing, fiction and non-fiction while section 4.3.2 compares the domain of financial news with general language and the varieties of fiction and non-fiction. The comparison is based on lexical sentiment feature distribution in the different corpora, providing an overview of both sentiment usage and polarity. The null hypothesis set out in equation 3 in these analyses is that the language of emotion, represented by lexical sentiment features, has the same distribution in two corpora representing different varieties, domains or special languages of English:

$$H_0 : \pi SentLang_{corpus1} = \pi SentLang_{corpus2} \tag{3}$$

The distributions of lexical sentiment features in the corpora are compared using the χ^2 test for independence to determine a statistical basis for rejecting the null hypothesis that the proportions of these features in the different corpora are the same. The alternative hypothesis is that usage patterns of affective text in these sub-languages represented by the corpora are statistically distinct which would justify a re-assessment of lexicon selection and use on an application– and domain–specific basis, or indeed the need for domain–specific lexica. While this analysis does not address domain–specific polarity values for individual lexical items, it does provides a strong empirical basis for work in the area of domain–specific sentiment analysis (Choi and Cardie 2009; Choi et al. 2009) as findings would strongly suggest that both the amount and the polarity bias of sentiment expressions are domain dependent in English.

4.3.1 Language Varieties: Fiction vs Non-Fiction

In order to investigate the distributions of affective text in language varieties of English, two corpora were selected for comparison: the imaginative and informative sub-corpora of the written text component of the BNC. The imaginative section contains mostly fiction but also other literary texts such

as poetry. The informative section consists of non-fictional expository writing mainly drawn from published books and periodicals. The intuition here might be that fictional writing would tend to be more affective than plain informative text and indeed this is what was found consistently. Although some features show no significant difference between the two corpora (GI_{sent} NegAff and Transaction Gain) and some are somewhat more common in the information corpus (GI_{sent} Fall and Positiv and WNA emotion–eliciting situation), overall sentiment features occur more often in imaginative text (fiction) than in informative text (non-fiction). This difference is statistically significant in a χ^2 test for independence at $p < 0.005$ for 26 features while two, WNA attitude and psychological response, show a statistically significant difference in proportions at $p < 0.05$ and $p < 0.01$ respectively. The results strongly suggest that the expression of sentiment is more common in the Imaginative corpus than in the Informative corpus.

Having determined that sentiment categories have genre-specific distributions, this second analysis explores whether sentiment polarity is also genre-specific. For both corpora and both the GI_{sent} and SWN lexica, positivity is dominant, as in general language but more pronounced in the sub-corpora than in general language, as set out in table 17. This extreme positive bias is most pronounced in the Informative corpus (GI_{sent} ratio 1:0.41; SWN ratio 1:0.76). In summary, although the Imaginative corpus contains more affective text than the Informative corpus, the affect expressed in fictional texts is less skewed overall towards positivity than in non-fictional writing where affective text content is less in volume but more biased towards the positive end of the affective spectrum. These findings would suggest that sentiment polarity bias is skewed in sub-varieties of English relative to general language, with the degree of skew dependent on variety type. An affective or polarity baseline for text therefore appears to be dependent on language variety. While this corpus analysis does not address the effect on readers of a polarity bias in text, it does highlight the need to investigate how these differences in distributions and polarity across language varieties are interpreted by readers and whether readers are sensitive to expected sentiment baselines for particular genres or varieties of text. These questions have major implications for the development of automatic sentiment identification systems and their adaptation to different language varieties.

Table 17 Positive and Negative Feature Counts and Ratios for Corpora

	Full BNC	**Imaginative**	**Informative**	**Finance**
GI_{sent} **Negative**	28039	9731	8375	25156
GI_{sent} **Positive**	43769	19608	20412	44860
GI_{sent} **Pos:Neg ratio**	**1 : 0.64**	**1 : 0.5**	**1 : 0.41**	**1 : 0.56**
SWN Neg	188121	91969	65105	160209
SWN Pos	198946	107181	85218	169956
SWN Pos:Neg ratio	**1 : 0.95**	**1 : 0.86**	**1 : 0.76**	**1 : 0.94**

4.3.2 Specialised Language: Financial News

In addition to this variety distinction, there is the issue of possible idiosyncracies of sentiment term usage in domain or specialised languages of English. The domain of financial news was selected for analysis here as the effect of news and news polarity on the financial markets has been the subject of intensive research in the domain of finance for a number of years. Engle and Ng (1993), for example, propose the asymmetric news impact curve which posits that negative news has a stronger and more long-lasting impact on market variables, in particular market volatility, than positive news. In this analysis, market variables, such as price movements are taken as a proxy for "news", with an unexpected price increase or decrease constituting "good" or "bad" news respectively. In more recent analyses, such as (Tetlock 2007), the text of the news itself has been used to generate a negative or positive sentiment index and again there is a statistically significant effect of "good" or "bad" news on market variables. Indeed, the topic has been absorbed into the financial mainstream with many financial software and content providers now offering SA add-ons for news feeds which claim to monitor sentiment as derived from news and the markets. Given the importance of sentiment and news sentiment indicators in the world of finance, this comparative corpus analysis of affective text distributions aims to examine whether financial language differs significantly from general language or varieties of English in terms of its use of affective terms. This investigation has potentially serious implications for the world of finance as research suggests that sentiment in financial news could be:

- a potential predictor of market movements;
- a potential cause of market movements;
- even a possible means of manipulating market movements.

If the usage of affective terms is statistically distinct in financial news as opposed to general language or other language varieties, the nature of these differences should be explored and it may even be necessary for financial regulators to control or at least monitor affective content of financial news and its effects on the markets. In addition to basic affective text usage, it is important to investigate any bias of financial news on the polarity spectrum relative to general language, as Engle and Ng (1993) posit that it is news *polarity* in particular which affects the markets. Again, the notion of polarity baselines and reader expectations given these baselines is highly relevant for the development of automated sentiment identification systems for finance. If different language varieties and domains have their own polarity bias, individual sentiment values in isolation are no longer informative, what becomes important are polarity values relative to context and expectations built up over time for a given domain.

To explore these issues, a corpus of approximately 2 million words of financial news was collected from news sources such as the financial sections of Reuters, Bloomberg, CNN, and various British, Irish and other national-

ity newspaper sources. The corpus was obtained automatically from Internet sources of these media and stripped of all mark-up. There are a total of 5633 written, non-fictional texts with an average of 355 words per text and an average of approximately 20 words per sentence. For the purposes of this analysis, the financial corpus was transformed into a frequency list identical in format to the WFWSE BNC list, including lemmas identified using the morpha tool from the University of Sussex (Minnen et al. 2001) and part–of–speech tags derived using the LT-POS tagger from the University of Edinburgh. This lemmatised and tagged corpus is compared with the BNC and its sub-corpora of imaginative and informative texts. Again, the proportion of sentiment features in each corpus is compared using the χ^2 test for independence in order to determine whether there is a statistical basis to reject the null hypothesis in 4: that the proportions of sentiment features used are the same across corpora.

$$H_0 : \pi_{finCorpus} = \pi_{corpus2} \tag{4}$$

Positive to negative polarity ratios are also compared to detect potential polarity bias specific to financial news texts.

The key finding is that there is a statistically significant difference in proportions of sentiment features between the financial news corpus and the full BNC, the Imaginative and the Informative sub-corpora. Financial news can be said to constitute a specialised language in its own right with regard to its affective term usage. While the proportion differences are statistically significant across the three corpus pairs, the differences are perhaps most stark with respect to the language variety sub-corpora where the results follow a definite trend towards higher frequency of affective terms in the financial news corpus. The results for comparison with the full BNC are more dependent on lexicon or feature type. The following sections set out the results for the comparison of sentiment feature use and polarity orientation in the three corpus pairs.

Financial Corpus and General Language (BNC). In a comparison of financial news with general language as represented by the BNC, the proportions of sentiment feature usage in the two corpora were found to be statistically significantly different for all sentiment features for which there were observations. However, there is no single trend towards greater or lesser frequency of sentiment expression overall in one or other corpus. Rather the difference in proportions is dependent on lexicon or feature type. SWN features are more prevalent in the BNC than the finance Corpus. This could be an artefact of the relative size and nature of the finance corpus. The finance corpus is 50 times smaller than the BNC and consists of financial news which constitutes a special language of English and as such may have a restricted vocabulary which avoids rare term use. This highlights a possible disadvantage of using very broad coverage resource, such as SWN, for domain specific applications where a more limited term set with higher frequency might be sufficient or indeed more appropriate. As regards the DAL, the Evaluation (good-bad) features are more prevalent in the BNC, whereas activation (representing the strong–weak dimension) have a stronger presence in the financial corpus. The

relative importance of the activation emotional dimension would suggest that strength and weakness are key factors in representing and interpreting financial news while in general language the evaluation good-bad dimension alone is much more dominant. This again highlights the need to assess domain–specific inclusion of sentiment features in any automated sentiment analysis system through the use of appropriate lexical or other resources.

All of the features of the WN Affect lexicon for which there are observations are statistically more frequent in the BNC corpus than in the finance corpus, some features over twice as frequent. This could be due to the nature of the WNA lexicon which aims to provide a lexicon of aspects of emotional experience rather than focusing solely on affective dimensions of terms. It is possible that financial news does not commonly refer to emotional experiences, rather it provides an affective interpretation of financial events. This could bring into question the utility of non-polarity lexica in financial sentiment analysis.

The distribution of GI_{sent} sentiment features between the two corpora, set out in table 18, is somewhat more complex. Unsurprisingly, transaction positive and negative features and the Fall feature are more prevalent in the financial corpus where much of the news reports on transactions and movement (of prices, shares, etc). Table 18 might suggest that the financial corpus tends towards more positive and domain–specific features with higher proportions of negative features in the BNC relative to the financial corpus. Indeed, the ratio of positive to negative features, shown in table 17, does show a stronger bias towards positivity in the financial corpus. According to a χ^2 test for independence, this difference in proportions of positive to negative features is statistically significant for GI_{sent} values ($\chi^2 = 147.06$, $df = 1, p < 0.0001$) but not for SWN values ($\chi^2 = 0.431$).

Table 18 GI_{sent} Sentiment Feature Proportion Dominance

More in Finance Corpus	**More in BNC**
Fail, Fall, Positive, TrnGain, TrnLoss	Hostile, NegAff, Negativ, Pain, PosAff, Vice, Wlbloss

Financial Corpus and Language Varieties: Fiction and Non-Fiction. Financial news has proven to be quite distinct from general language but does it conform more to the language of fictional or non-fictional writing, as represented by the BNC sub-corpora. Overall, the financial corpus shows much higher frequency of almost all sentiment features than the sub-corpora. Only the WNA features have some features which are more prominent in the sub-corpora. The division in WNA features follows somewhat the characteristics of the long–term:positive / short–term:negative distinction noted in section 3.3.3 with "long–term", positive features (Attitude and Trait) more frequent in the financial corpus and "short–term", negative features (Mood and Responses) more frequent in both the Imaginative and Informative corpora. While the

feature distributions are similar for the two sub-corpora, the degree to which they differ from the financial corpus differs. The language of finance is a closer approximation to fictional writing, with sentiment features 1–3 times more frequent than in the imaginative corpus but 3-5 times more frequent than in the informative corpus. Furthermore, the polarity bias of financial news has a weaker positive bias than imaginative and much weaker than informative writing, as illustrated in table 17. Financial news in fact stands as a half-way point between general language and fiction - more positively biased than general language, less biased towards positive than fiction and much less biased than informative text.

Implications for Sentiment Analysis. This comparative corpus analysis has identified some of the key affective characteristics of financial text with respect to English in general and some of its varieties. Firstly, affective text usage is very frequent in financial news. Secondly, both the evaluation and activation dimensions of emotion are prominent in the financial corpus. Thirdly, financial news has the status of a specialised language of English with contingent restricted lexical choices. Finally, the positive polarity bias of financial news is statistically significant and distinct from the bias of general language and language varieties. Financial news appears to be marginally more positive than general language and marginally less than the two language varieties investigated. These domain–specific characteristics have strong implications for sentiment analysis in general and in finance. For sentiment analysis applications in finance, it would be useful to represent both evaluation and activation dimensions of emotion as these two features are highly frequent in financial news. For any domain-specific application, it may not be necessary to use a broad coverage lexicon as, in addition to domain–specific semantic variation within lexical items, the lexicon of the domain itself may be restricted and a domain–specific lexicon the optimal solution for SA. For any sentiment analysis application, the sentiment value derived from any text must be interpreted in the context of some baseline polarity metric for the relevant language domain or variety. Sentiment polarity is not homogeneous across language varieties and this baseline represents reader expectations and assumptions and it is only in this context that a polarity value can have meaning. Borrowing from econometrics, this baseline could be represented as a time series of polarity values and it is changes or volatility in a polarity series which become important, not raw values. This case study highlights some generic requirements of sentiment analysis systems but also the need to evaluate any application domain thoroughly in order to estimate any domain–specific idiosyncracies which must be addressed.

5 Conclusions and Future Work

This paper has detailed a comprehensive analysis of four lexical resources for sentiment analysis in common usage today. The lexical content and sentiment

feature assignment of each lexicon has been evaluated, individually and in relation to each other. The results of this analysis showed that the lexical resources are consistent with each other in terms of their sentiment feature assignments and their lexical content. This finding in a sense validates the lexica in so far as each has very different origins, theoretical underpinnings and development criteria yet what they represent and how they represent it remains largely consistent across lexica. However, although they are consistent in many respects, there is sufficient difference between the four resources, in terms of content, representation and coverage, to merit careful consideration of individual characteristics for possible impacts on an automated sentiment analysis system.

Corpus analysis confirms that the sentiment lexica in combination constitute a distinct sub–set of the English language with characteristics which are statistically distinct from general language. The distribution of terms and features for each lexicon in English has been evaluated relative to a general language corpus (the BNC), two language variety corpora (BNC Imaginative and Informative corpora) and one special language corpus of financial news texts. The results strongly suggest that affective text content is not homogeneous across different language varieties or domains of use. Furthermore, results would indicate that the polarity of sentiment in text in general tends to be asymmetric with a positive skew. This bias, however, is also not homogeneous across language varieties. Although in this analysis the direction of the bias does not change between corpora, there is a statistically significant difference in intensity of bias between corpora.

The findings suggest that, although there does appear to be a language of sentiment distinct from general language, there is not one size that fits all in terms of degree and range of sentiment expression across language varieties and domains. As affective text content and polarity appear to be dependent on language variety, the notion of a polarity baseline for a given domain against which an automated sentiment analysis system can evaluate its results becomes essential. Econometric analysis suggests that indeed people are sensitive to and form expectations regarding the polarity of news in the financial domain at least. In anecdotal evidence from studies we have carried out with human annotators, participants often comment on the sensitivity of their responses to negative elements, even very small elements, in text, particularly at positions of prominence such as at the start or end of a text. A key avenue for future research in sentiment analysis is to determine whether people are sensitive to this polarity baseline and how they react to violations of their expectations in this regard. It is the parameters of such reactions to changes in a polarity baseline and the domain–specific nature of this baseline which we aim to determine in future work.

References

Khurshid Ahmad, David Cheng, and Yousif Almas. Multilingual sentiment analysis in financial news streams. In Stefano Cozzini, Stefano d'Addona, and Rosario Mantegna,

editors, *Proc. of the 1st Intl Conf on Grid in Finance*, Palermo, Italy, 2006.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*, 2010.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*, 2007.

Sergio Bolasco and Francesca della Ratta-Rinaldi. Experiments on semantic categorization of texts: analysis of positive and negative dimension. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data (JADT 2004)*, 2004.

Jerry Boucher and Charles E. Osgood. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, 8(1):1–8, 1969.

Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 590–598, Singapore, 2009.

Yoonjung Choi, Youngho Kim, and Sung-Hyon Myaeng. Domain-specific sentiment analysis using contextual feature generation. In *TSA '09 Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44. ACM New York, 2009.

Ann Devitt and Khurshid Ahmad. Cohesion-based sentiment polarity identification in financial news. In *Proceedings of ACL 2007*, Prague, Czech Republic, 2007.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

Paul Ekman. Strong evidence for universals in facial expressions: A reply to russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994.

Paul Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129, 1971.

Robert F. Engle and Victor K. Ng. Measuring and testing the impact of news on volatility. *Journal of Finance*, 48(5):1749–1778, 1993.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*, Genova, Italy, may 2006.

N. L. Etcoff and J. J. Magee. Categorical perception of facial expressions. *Cognition*, 44: 227–240, 1992.

Christiane Fellbaum. *WordNet,an electronic lexical database*. The MIT Press, Cambridge, Mass., 1998.

Vasileios Hatzivassiloglou and Kathy R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, pages 174–181, New Brunswick, NJ, 1997.

Vasileios Hatzivassiloglou and J. M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics*, New Brunswick, NJ, 2000.

Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland, 2004.

Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of COLING 2004*, Geneva, 2004.

Harold D. Lasswell and Abraham Kaplan. *Power and society : a framework for political inquiry*. Yale University Press, New Haven, 1950.

Geoffrey Leech, Paul Rayson, and Andrew Wilson. *Word Frequencies in Written and Spoken English*. Longman, London, 2001.

J.R. Martin and P.R.R. White. *Language of Evaluation: Appraisal in English*. Palgrave Macmillan, 2005.

Rada Mihalcea and Hugo Liu. A corpus-based approach to finding happiness. In *Proceedings of AAAI Spring Symposium on Computational Approaches to Weblogs*, Menlo Park,

California, 2006. AAAI Press.
G. Minnen, J. Carroll, and D. Pearce. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223, 2001.
Mark L. Mitchell and J. Harold Mulherin. The impact of public information on the stock market. *Journal of Finance*, 49(3):923–950, 1994.
J. Zvi Namenwirth and Robert Philip Weber. *Dynamics of Culture*. Allen and Unwin, Boston, Massachussetts, 1987.
Tetsuya Nasukawa and Jeonghee Yi. Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003)*, Sanibel Island, Florida, USA, 23-25 October 2003.
Paula M. Niedenthal and Jamin B. Halberstadt. Emotional response as conceptual coherence. In Eric Eich, John F. Kihlstrom, Gordon H. Bower, Joseph P. Forgas, and Paula M. Niedenthal, editors, *Cognition and Emotion*, chapter 4, pages 169–203. Oxford University Press, Oxford, 2000.
Andrew Ortony, Gerard L. Clore, and Mark A. Foss. The referential structure of the mental lexicon. *Cognitive Science*, 11(3):341–364, 1987.
C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of meaning*. University of Illinois Press, Chicago, Ill, 1957.
Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL 2004*, pages 271–278, Barcelona, Spain, 2004. Association of Computational Linguistics.
Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP'02*, pages 79–86, 2002.
Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Proceedings of the AAAI Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
James A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994.
James A. Russell and Albert Mehabrian. Evidence of a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 1977.
P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. M.I.T Press, Cambridge, 1966.
Carlo Strappavara and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *Proceedings of LREC 2004*, Lisbon, Portugal, 2004.
Kevin Sweeney and Cynthia Whissell. A dictionary of affect in language: I. establishment and preliminary validation. *Perceptual and Motor Skills*, 1984.
Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168, 2007.
E. L. Thorndike and Irving Lorge. *The Teacher's Word Book of 30,000 Words*. Teachers College, Columbia University, 1944.
Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL'02*, pages 417–424, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.
Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21 (4):315–346, 2003.
David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235, 1985.
Cynthia Whissell. The dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *Emotion: theory research and experience*, volume 4, The measurement of emotions. Academic Press, London, 1989.
Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. Recognizing and organizing opinions expressed in the world press. In *AAAI Spring Symposium on New Directions in Question Answering*, Stanford University, Stanford, CA, USA, 2003. AAAI Press.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, sept 2004.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural LanguageProcessing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, oct 2005. Association for Computational Linguistics.