Using Appraisal Groups for Sentiment Analysis

Casey Whitelaw
Language Technologies
Research Group
School of Information
Technologies
University of Sydney
Sydney, NSW, Australia
casey@it.usyd.edu.au

Navendu Garg
Linguistic Cognition Lab
Dept. of Computer Science
Illinois Institute of Technology
10 W. 31st Street
Chicago, IL 60616, USA
gargnav@iit.edu

Shlomo Argamon
Linguistic Cognition Lab
Dept. of Computer Science
Illinois Institute of Technology
10 W. 31st Street
Chicago, IL 60616, USA
argamon@iit.edu

ABSTRACT

Little work to date in sentiment analysis (classifying texts by 'positive' or 'negative' orientation) has attempted to use fine-grained semantic distinctions in features used for classification. We present a new method for sentiment classification based on extracting and analyzing *appraisal groups* such as "very good" or "not terribly funny". An appraisal group is represented as a set of attribute values in several task-independent semantic taxonomies, based on Appraisal Theory. Semi-automated methods were used to build a lexicon of appraising adjectives and their modifiers. We classify movie reviews using features based upon these taxonomies combined with standard "bag-of-words" features, and report state-of-the-art accuracy of 90.2%. In addition, we find that some types of appraisal appear to be more significant for sentiment classification than others.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms,Experimentation

Keywords

Opinion Mining, Sentiment Analysis, Text Classification, Shallow Parsing, Review Classification, Appraisal Theory

1. INTRODUCTION

Recent years have seen a growing interest in *non-topical* text analysis, in which characterizations are sought of the opinions, feelings, and attitudes expressed in a text, rather than just the facts. The recent AAAI Spring Symposium on *Exploring Attitude and Affect in Text* [17], with over 60 attendees, reflects the growing importance of this area of research. A key problem in this area is *sentiment classification*, in which a document is labelled as a positive ('thumbs up') or negative ('thumbs down') evaluation of a target object (film, book, product, etc.). Immediate applications include data and web mining, market research, and customer relationship management.

A primary testbed task for sentiment classification has been the classification of movie reviews. Reviews offer an interesting and difficult test case for sentiment analysis. Opinions are expressed in many complex ways (including sarcasm and metaphor), and there is much unrelated and potentially misleading text such as plot synopses.

To date, most work on sentiment analysis has relied on two main approaches. The first ("bag of words") attempts to learn a positive/negative document classifier based on occurrence frequencies of the various words in the document; within this approach various learning methods can be used to select or weight different parts of a text to be used in classification. The other main approach ("semantic orientation") classifies words (usually automatically) into two classes, "good" and "bad", and then computes an overall good/bad score for the text.

However, such approaches miss important aspects of the task. First, a more detailed semantic analysis of attitude expressions is needed, in the form of a well-designed taxonomy of attitude types and other semantic properties (as noted by Taboada and Grieve [18]). Second, the "atomic units" of such expressions are not individual words, but rather *appraisal groups*: coherent groups of words that express together a particular attitude, such as "extremely boring", or "not really very good".

This paper addresses both of these issues directly, by focusing on the extraction and analysis of *adjectival appraisal groups* headed by an appraising adjective (such as 'beautiful' or 'boring') and optionally modified by a sequence of modifiers (such as 'very', 'sort of', or 'not'). We have adopted taxonomies for the attributes of such expressions from Martin and White's Appraisal Theory [10], developed within the tradition of Systemic Functional Linguistics [6]. We built a lexicon using semi-automatic techniques, gathering and classifying 1329 adjectives and modifiers to categories in several taxonomies of appraisal attributes. We heuristically extract adjectival appraisal groups from texts and compute their attribute values according to this lexicon. Documents were

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
Copyright 2005 ACM 1-59593-140-6/05/0010 ...$5.00.

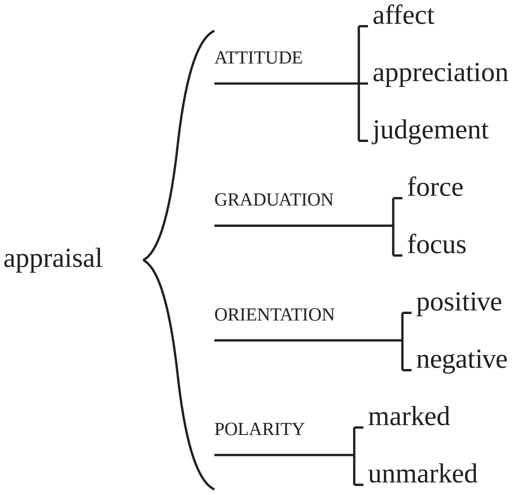

Figure 1: Main attributes of Appraisal **and their highest-level options.**

then represented as vectors of relative frequency features computed over these groups and a support vector machine learning algorithm [3] was used to learn a classifier discriminating positively from negatively oriented test documents.

We have applied this approach to movie review classification with positive results. Despite the low coverage of our current lexicon, adjectival appraisal group features alone give decent classification performance (78%). When coverage is improved by the simple expedient of adding in simple bag of words features, classification accuracy reaches 90%, higher than previously published results.

2. APPRAISAL GROUPS

Appraisal denotes how language is used to adopt or express an attitude of some kind towards some target. For example, in "I found the movie quite monotonous", the speaker (the *Appraiser*) adopts a negative *Attitude* ("monotonous") towards "the movie" (the *Appraised*). Note that attitudes come in different types; for example, 'monotonous' describes an inherent quality of the Appraised, while 'loathed' would describe the emotional reaction of the Appraiser.

A full *appraisal expression* is thus a piece of text (usually a clause, but possibly larger) expressing appraisal of some sort. Appraisal expressions are the basic 'atoms' for analysis of how attitudes are expressed in a text, and so extracting them is the basic task for appraisal analysis. By analogy to information extraction, we consider representing an appraisal expression as a frame filled with several slot values, giving (at least) the Appraiser, Appraised, Appraisal Type, and Orientation (positive/negative). For example, appraisal in the sentence

> I truly believe that this is the best film I've seen this year.

could be represented by the frame

$$\begin{pmatrix} \text{Appraiser:} & \text{writer} \\ \text{Appraised:} & \text{this (film)} \\ \text{Attitude:} & \text{appreciation/reaction-quality} \\ \text{Orientation:} & \text{positive} \end{pmatrix}$$

Clearly, extracting all of this information accurately is a difficult process, requiring identifying who is talking about what in potentially complex texts (not to mention the need for coreference resolution, which is far from solved). We therefore first address a useful but simpler problem, that of extracting *appraisal groups*, defined as those groups and phrases in a text giving what kind and intensity of appraisal is expressed. We consider in this paper extraction of a main type of appraisal group, *adjectival* appraisal groups, which give good classification results despite seemingly low coverage in the corpus. We thus expect future inclusion of nominal ("a total mess") and verbal ("absolutely loved") appraisal groups to further improve results.

2.1 Taxonomies of appraisal

Our first goal is to extract appraisal groups, from which we then derive useful features for machine learning. Following Martin and White [10], we will assign four main types of attributes (Fig. 1) to appraisal groups: Attitude, Orientation, Graduation, and Polarity[1]:

Attitude gives the type of appraisal being expressed as either *affect*, *appreciation*, or *judgement*. Affect refers to a personal emotional state (e.g., 'happy', 'angry'), and is the most explicitly subjective type of appraisal. The other two options express evaluation of external entities, differentiating between evaluation of intrinsic *appreciation* of object properties (e.g., 'slender', 'ugly') and social *judgement* (e.g., 'heroic', 'idiotic'). Figure 2 gives a more detailed view of the various options in Attitude, together with illustrative adjectives. In general, attitude may be expressed through nouns (e.g., 'triumph', 'catastrophe') and verbs (e.g., 'love', 'hate'), as well as adjectives.

Orientation is whether the appraisal is *positive* or *negative* (often simply termed 'sentiment').

Graduation describes the intensity of appraisal in terms of two independent dimensions of *force* (or 'intensity') and *focus* ('prototypicality'). Graduation is largely expressed via modifiers such as 'very' (increased force), 'slightly' (decreased force), 'truly' (sharpened focus), or 'sort of' (softened focus), but may also be expressed lexically in a head adjective, e.g., 'greatest' vs. 'great' vs. 'good'.

Polarity of an appraisal is *marked* if it is scoped in a polarity marker (such as 'not'), or *unmarked* otherwise. Other attributes of appraisal are affected by negation; for example, "not good" expresses a different sentiment from "good".

From this perspective, most previous sentiment classification research has focused exclusively upon Orientation, with Attitude type addressed only indirectly, through the use of bag-of-words features. An exception is Taboada and Grieve's [18] method of automatically determining top-level attitude types via application of of Turney's PMI method [19]. They observed that different types of reviews contain different amounts of each attitude-type.

[1] Note we use the term 'Polarity', as in SFL, to denote the grammatical notion "explicit negation of a quality or assertion within the scope of the particle 'not' or the equivalent"; note that this term has also been used in the literature to mean what we refer to as Orientation.

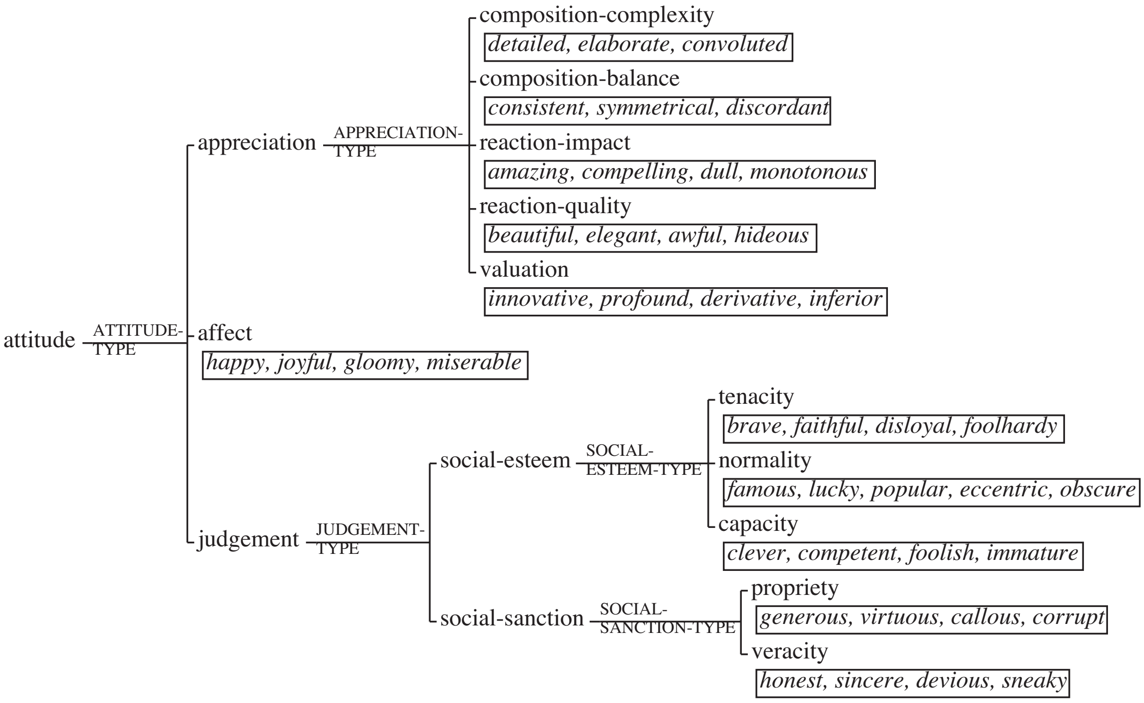
attitude
ATTITUDE-TYPE
appreciation
APPRECIATION-TYPE
composition-complexity
detailed, elaborate, convoluted
composition-balance
consistent, symmetrical, discordant
reaction-impact
amazing, compelling, dull, monotonous
reaction-quality
beautiful, elegant, awful, hideous
valuation
innovative, profound, derivative, inferior
affect
happy, joyful, gloomy, miserable
judgement
JUDGEMENT-TYPE
social-esteem
SOCIAL-ESTEEM-TYPE
tenacity
brave, faithful, disloyal, foolhardy
normality
famous, lucky, popular, eccentric, obscure
capacity
clever, competent, foolish, immature
social-sanction
SOCIAL-SANCTION-TYPE
propriety
generous, virtuous, callous, corrupt
veracity
honest, sincere, devious, sneaky

Figure 2: Options in the Attitude taxonomy, with examples of appraisal adjectives from our lexicon.

$$\underset{\text{“happy”}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \text{Orientation:} & \text{positive} \\ \text{Force:} & \text{neutral} \\ \text{Focus:} & \text{neutral} \\ \text{Polarity:} & \text{unmarked} \end{pmatrix}} \Rightarrow \underset{\text{“very happy”}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \text{Orientation:} & \text{positive} \\ \textbf{Force:} & \textbf{high} \\ \text{Focus:} & \text{neutral} \\ \text{Polarity:} & \text{unmarked} \end{pmatrix}} \Rightarrow \underset{\text{“not very happy”}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \textbf{Orientation:} & \textbf{negative} \\ \textbf{Force:} & \textbf{low} \\ \text{Focus:} & \text{neutral} \\ \textbf{Polarity:} & \textbf{marked} \end{pmatrix}}$$

Figure 3: Analysis of appraisal group “not very happy”.

An *appraisal group* (in English) comprises a *head adjective* with defined attitude type, with an optional preceding list of *appraisal modifiers*, each denoting a transformation of one or more appraisal attributes of the head. For example, “not extremely brilliant”, has head ‘brilliant’ and modifiers ‘not’ and ‘extremely’. We take advantage of typical English word-ordering and use all pre-modifiers, allowing for intervening articles and adverbs. This allows groups such as “not *all that* good” or “truly *a* really horrible”, where ‘not’ and ‘truly’ modify ‘good’ and ‘horrible’, respectively. We treat modifiers as having nested scope, so that transformations to appraisal attributes are applied inside out[2]. Figure 3 shows the derivation of the appraisal attributes of “not very happy”. Derivation relies on a suitable lexicon of appraisal attributes, which we describe below.

2.2 The lexicon

We used a semi-automated technique to construct a lexicon giving appraisal attribute values for relevant terms. A value for each appraisal attribute is stored for each appraisal adjective[3]; for example, the lexical entry for ‘beautiful’ reads:

$$\left[\begin{array}{ll} \text{‘beautiful’} & \\ \text{Attitude:} & \text{appreciation/reaction-quality} \\ \text{Orientation:} & \text{positive} \\ \text{Force:} & \text{neutral} \\ \text{Focus:} & \text{neutral} \\ \text{Polarity:} & \text{unmarked} \end{array}\right.$$

Modifiers, mostly adverbs, give transformations for one or more appraisal attributes, for example:

$$\left[\begin{array}{ll} \text{‘very’} & \\ \text{Force:} & \text{increase} \end{array}\right.$$

or polarity modification:

$$\left[\begin{array}{ll} \text{‘not’} & \\ \text{Orientation:} & \text{negate} \\ \text{Force:} & \text{reverse} \\ \text{Polarity:} & \text{marked} \end{array}\right.$$

[2] While there are certainly cases where such nesting is inaccurate, as in “not even good” where “not even” should be treated as a nested unit, we find that assuming right-nesting is a reasonable approximation for the time being.

[3] Force and Focus are both ‘neutral’ by default. Comparative (JJR) and superlative (JJS) adjectives are assigned ‘high’ and ‘maximum’ force respectively, though other lexical gradations of intensity (e.g., ‘love’ vs ‘like’) are not currently addressed.

seeds	12	'arresting', 'captivating', 'engaging', 'fascinating', 'exciting', 'moving', ...
candidates	435	'striking' (5), 'noteworthy' (5), 'impressive' (5), 'thrilling' (4), ... 'industrious' (1), 'executive' (1), 'peculiar' (1), 'magnetic' (1)
selected	61	'striking', 'impressive', 'thrilling', 'stirring', 'enchanting', 'breathtaking', 'mind-blowing', ...

Figure 4: Example of seed term expansion and manual selection, for Appreciation/Reaction-Impact:Positive.

Modifiers can affect multiple appraisal attributes at once; e.g., 'really' functions both as an intensifier of force and a sharpener of focus.

To build the lexicon, we started with the example words and phrases given for various appraisal options in [10] and [11] as seed terms, using a semi-automated technique to quickly build a lexicon with decent coverage of adjectival appraisal. Modifier seed terms were generated similarly, by finding adverbs collocating with adjective seed terms in our corpus. Candidate expansions for each seed term were generated from WordNet and from two online thesauri[4]. In WordNet, the members of each synset were taken as the related set; similarly, synonym and related word lists were taken from each thesaurus. Candidates were accepted only with the same part of speech as a seed term.

Candidate lists for all terms in one category were pooled and all candidate terms ranked by frequency of occurrence in the various candidate lists. This provided a coarse ranking of relevance, enabling more efficient manual selection. Uncommon words, unrelated words, or words arising from an incorrect sense of the seed term will tend to occur less frequently in the candidate list than those that are related to more of the seed terms and are present in more of the resources. As well as increasing coverage, using multiple thesauri allows for more confidence votes and in practice increases the utility of the ranking.

Each ranked list was manually inspected to produce the final set of terms used. In practice, terms with low confidence were automatically discarded, reducing the amount of manual work required. Figure 4 shows a partial example of the expansion and selection process. We note that some element of human subjectivity is inherent in this process; future work will include improving standardization of the lexicon by involving multiple lexicon builders and evaluating inter-rater reliability of attribute assignment.

In total, 1329 terms were produced from 400 seed terms, in around twenty man-hours. This lexicon is by no means complete, but is large enough to investigate the usefulness of this approach. Appraisal adjectives in the lexicon cover 29.2% of adjectives in the testbed corpus, comprising 2.7% of words in the corpus. Note that the appraisal taxonomies used in this work are general purpose, and were not developed specifically for sentiment analysis or movie review classification. Thus we expect appraisal group analysis to be highly portable to other related tasks.

3. FEATURE SETS

The standard approach to representing documents as multidimensional vectors as input for machine learning techniques is to measure the frequency of various text elements relative to the total number of such elements (words, e.g.) in the text. We follow that method here as well, defining features as various disjunctions of lexical items or appraisal group attribute values as defined in our appraisal taxonomies. Raw counts are thus normalized against the total number of units of the corresponding type in the text[5]. This gave us the following feature sets:

W:A *Words by Attitude* — Frequency of each adjective with a defined attitude type, normalized by total number of such adjectives in the text.

S:A *Systems by Attitude* — Total frequency of attitude adjectives for each Attitude option (at every level in the taxonomy), normalized by total number of such adjectives in the text.

S:AO *Systems by Attitude and Orientation* — Total frequency of attitude adjectives for each combination of Attitude and Orientation (e.g., Orientation=*positive* and Attitude=*affect*), normalized by total number of such adjectives in the text.

G:A *Appraisal Group by Attitude* — Total frequency of appraisal groups with each possible Attitude, normalized by total number of appraisal groups in the text.

G:AO *Appraisal Group by Attitude & Orientation* — Total frequency of appraisal groups with each possible combination of Attitude and Orientation, normalized by total number of appraisal groups in the text.

G:AOF *Appraisal Group by Attitude, Orientation, & Force* — Total frequency of appraisal groups with each possible combination of Attitude, Orientation, and Force, normalized by total number of appraisal groups in the text.

BoW *Bag of Words* — relative frequencies of all words in the text.

BoW+G:AO Union of BoW and G:AO.

BoW+G:AOF Union of BoW and G:AOF.

The next section describes our results for sentiment classification using these various feature sets.

4. EXPERIMENTS

4.1 Methodology

To test the usefulness of adjectival appraisal groups for sentiment analysis, we evaluated the effectiveness of the above feature sets for movie review classification, using the publicly available collection of movie reviews constructed by Pang and Lee [14]. This standard testbed consists of 1000 positive and 1000 negative reviews, taken from the IMDb movie review archives[6]. Reviews with 'neutral' scores (such as three stars out of five) were removed by Pang and Lee,

[4] `http://m-w.com` and `http://thesaurus.com`

[5] In preliminary experiments, the use of relative frequencies within each node of the taxonomy, as in [1, 20], gave inferior results to this simpler procedure.

[6] See `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

Feature Set	N_{feat}	CV Acc.	RS Acc.
W:A	1047	77.6	77.7
S:A	1355	78.0	77.8
S:AO	1278	78.2	77.4
G:A	1136	78.2	77.8
G:AO	1597	78.6	78.0
G:AOF	2147	78.3	77.9
BOW	48,314	87.0	87.6
BOW+G:AO	49,911	90.2	90.1
BOW+G:AOF	50,461	88.7	89.6
P&L-04		*87.2*	
M&C-03(TVO)		*69.0*	
M&C-03(best)		*86.0*	

Table 1: Estimated mean accuracy results for different feature sets using 10-fold cross-validation (CV) and 40 randomized samples (RS); N_{feat} gives the total number of non-constant features in each feature set. For comparison, "P&L-04" denotes the best results obtained on this data set by Pang and Lee [14]; "M&C-03" denotes results obtained on the earlier movie review dataset by Mullen and Collier [12] for two feature sets.

giving a data set with only clearly positive and negative reviews[7]. Each document is preprocessed into individual sentences and decapitalized. We used an implementation of Brill's [2] part-of-speech tagger to help us find adjectives and modifiers.

Classification learning was done using WEKA's [23] implementation of the SMO [16] learning algorithm, using a linear kernel and the default parameters. It is possible that better results might be obtained by using a different kernel and tuning the parameter values, however such experiments would have little validity, due to the small size of the current testbed corpus.

4.2 Results

We evaluated sentiment classification accuracy using SMO with default parameters and a linear kernel. Evaluation was performed both using the standard 10-fold cross-validation, as well as using a sequence of 40 independently chosen random train-test partitions of the corpus with 1950 training and 50 test documents per run (enabling statistical testing). We trained models and evaluated test set accuracy for each of the feature sets described above. Mean accuracies for both cross-validation and random sampling are given in Table 1, with mean paired accuracy differences by random sampling in Table 2; significance was measured by two-tailed t-test.

For comparison, Table 1 also lists previous results from two previous studies. Directly comparable is the highest previous accuracy for this dataset, attained by Pang and Lee [14] via a complex combination of 'subjectivity clustering' and bag-of-words classification for sentiment analysis. We also show two results from Mullen and Collier [12], the only previous work we are aware of to use something akin to attitude type for sentiment analysis. Unfortunately, their results are not directly comparable with ours, as they used an earlier version of the movie review corpus (with only 1380 reviews); we show their results using only Turney Value and Semantic Differentiation features (TVO) as well as their best result for that corpus.

First, despite the low coverage of our lexicon, the baseline of using just attitude-bearing adjectives is reasonably high. This bears out our contention that attitude-bearing adjectives specifically are a key feature in the expression of sentiment. Using attitude type and orientation of these terms yields essentially the same accuracy, however. When using appraisal groups, which include the effect of appraisal modifiers, we see that using both Attitude Type and Orientation, we get a slightly higher accuracy (however, this is not significant). The small size of any increase is likely due to the fact that out of 41082 appraisal groups in the corpus, just 751 (1.8%) have their orientation flipped by marked polarity. While the inclusion of Force appears to bring no advantage here, this may be due to the current low granularity of Force and Focus distinctions in our lexicon.

Next, we note that all of the limited-coverage appraisal feature sets are outperformed by standard bag-of-words classification using all words (BoW), which attains a statistically significant 87.0/87.6% accuracy, competitive with Pang and Lee's [14] result on this dataset, based on classifying texts after extracting subjective passages from them. More significantly, we improve clearly on that result (attaining 90.2/90.1% accuracy) by combining appraisal group features (Attitude Type and Orientation) with the bag-of-words features (for coverage), demonstrating how appraisal analysis helps sentiment classification. This improvement is significant to a 99% confidence level (see Table 2). Again we note that including Force does not seem to help.

Space does not permit listing all the features for positive and negative documents, so we summarize our findings here. Of the 200 most significant features (100 for each of the positive and negative classes) in the model built from BoW+G:AO, 57 (28%) are systemic features. The vast majority of those (39) are drawn from subtypes of *appreciation*, with 14 (six positive and eight negative) from *judgement* and four (three positive and one negative) from affect. Appreciation thus appears to be the most central type of attitude for sentiment analysis (at least for movie review classification). In addition, while some adjectival features in BoW are included (duplicating work done by G:AO), many BoW features are clearly helping with coverage, including many nouns (e.g., 'mess', 'script', 'nothing', 'job', 'truth') and some verbs ('loved', 'wasted', 'delivered'), as well as other parts-of-speech[8].

5. RELATED WORK

An early, and still common, approach to sentiment analysis has been to use the so-called 'semantic orientation' (SO) of terms as the basis for classification [7]. Semantic orientation is derived from corpus-based collocation information; it is common to use search engines and treat the internet as a very large corpus, estimating SO for terms based on pointwise mutual information with certain anchor terms such as 'good' and 'poor' [19]. This is equivalent, in our approach, to using just Orientation values, computing a weighted sum

[7] This lack of 'inconclusive' documents may limit the real-world applicability of results on this dataset. An open research issue is how to deal effectively with such documents.

[8] Curiously, 'and', 'also', and 'as' are strong features for positive sentiment. This may indicate that rhetorical structure [9, 1] is also important for understanding sentiment.

	Comparison							
Base	S:A	S:AO	G:A	G:AO	G:AOF	BOW	BOW + G:AO	BOW + G:AOF
W:A	+0.05	-0.35	+0.05	+0.25	+0.02	+9.95**	+12.45**	+11.85**
S:A	–	-0.40	0.00	+0.20	+0.15	+9.90**	+12.40**	+11.80**
S:AO		–	+0.40	+0.60	+0.55	+10.30**	+12.80**	+12.20**
G:A			–	+0.20	+0.15	+9.90**	+12.40**	+11.80**
G:AO				–	-0.05	+9.7**	+12.20**	+11.60**
G:AOF					–	+9.75**	+12.25**	+11.65**
BOW						–	+2.5**	+1.90*
BOW+G:AO							–	-0.60

Table 2: Estimates of mean accuracy differences between all pairs of "Comparison" feature sets and "Base" feature sets. Statistical significance of the differences was tested using the two-sided t-test, as indicated: *=$p < 0.05$ and **=$p < 0.01$.

for the whole document.

Early attempts at classifying movie reviews used standard bag-of-words techniques with limited success [15]. The addition of typed features of semantic orientation has been shown to improve results [12]. Semantic orientation has also been useful for classifying more general product reviews [19]; that work has suggested that product reviews may be easier to classify than movie reviews, as they tend to be shorter and be more explicitly evaluative.

There have been some previous attempts at using a more structured linguistic analysis of text for sentiment classification, with mixed results. Mullen and Collier [12] produced features based on Osgood's Theory of Semantic Differentiation, using WordNet to judge the 'potency', 'activity', and 'evaluative' factors for adjectives. Using these features did not yield any reliable benefit, although it is unclear whether this is due to the theory or to its implementation. The same study used a small corpus of music reviews manually annotated for artist and work, and showed that knowing the Appraised can increase performance. Nasukawa and Yi [13] use POS-tagging, chunking, and manual templates to investigate 'sentiment verbs' and 'sentiment transfer verbs' and have shown that this approach can aid high-precision analysis. Previous work on including polarity ("good" vs. "not good") have given inconsistent results—either a slight improvement [15] or decrease [5] from bag-of-word baselines; our results show it to help slightly.

An additional problem (not addressed in this paper) facing sentiment classification is determining which parts of the text are relevant. Ideally, only subjective text that is expressed by the reviewer and deals directly with the item under review should be kept, but has so far proved difficult to isolate. Subjective use of verbs, adjectives and multi-word expressions can be learnt automatically and used to detect sentence-level subjectivity [21]. Adjectives play a strong role in subjective language, especially the class of 'gradable' adjectives [8] that can take modifiers such as 'very'.

More generally, Wilson et al. [22] have recently addressed learning models for finding opinion clauses and identifying their properties (mainly what we term force and orientation), based on clauses' lexical and syntactic properties. Using this approach, Pang and Lee [14] have applied a clustering approach to extract 'subjective' passages from texts. They show that classification learning applied to such extracts is more effective than classification based on the entire document.

6. DISCUSSION AND FUTURE WORK

We have shown that use of features based on appraisal group analysis can significantly improve sentiment classification, despite the low coverage of our current appraisal lexicon. Our results thus underscore the need to develop detailed and varied semantic tools to support sentiment analysis. In addition to improved accuracy, such taxonomic features can provide useful information about how language is used to express sentiment, as we observe above that one type of appraisal (*appreciation*) is more significant for classifying movie reviews. This type of insight is only enabled by a taxonomic analysis of appraisal type.

Our results show that even small hand-built ontologies can be useful, especially in combination with simpler high-coverage methods. Requiring a high level manual intervention is not ideal; due to the functional nature of the groupings, system networks would appear ripe for inference using current statistical thesaurus-building techniques [4]. This would also allow the construction of domain-specific ontologies, as opposed to the generic lexicon used in this paper.

Without some form of summarization or filtering, performance is necessarily limited by the presence of extraneous and potentially misleading appraisal in the document. We have made no attempt to include existing methods such as subjectivity summarization or position-based weighting, which we would expect to provide similar gains as have been seen in prior research.

We believe that the major challenge currently in sentiment analysis is the accurate identification of relevant full Appraisal Expressions including the Appraiser and Appraised in addition to appraisal type and orientation. This would enable more fine-grained analysis of the expressions of sentiment in a document. As applications requiring text classification grow to include far more than traditional topic-based tasks, there is a growing need for a more structured semantic approach to feature extraction and representation. Existing linguistic theories such as Appraisal Theory provide possible bases for new textual features which, as we have shown, can improve upon the results of traditional word-based techniques.

7. REFERENCES

[1] S. Argamon and J. T. Dodick. Conjunction and modal assessment in genre classification. In *AAAI Spring Symp. on Exploring Attitude and Affect in Text*, 2004.

[2] Eric Brill. A simple rule-based part of speech tagger. In *Proc. of ACL Conference on Applied Natural Language Processing*, Trento, Italy, 1992.

[3] N. Cristianini and J. Shaw-Taylor. *An Introduction to Support Vector Machines*. Cambridge Press, 2000.

[4] J. R. Curran. Ensemble methods for automatic thesaurus extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 222–229, Philadelphia, PA, USA, 2002.

[5] D. Dave and S. Lawrence. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. Twelfth Int'l World Wide Web Conference (WWW2003)*, 2003.

[6] Michael A. K. Halliday. *Introduction to Functional Grammar*. Edward Arnold, second edition, 1994.

[7] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In P. R. Cohen and W. Wahlster, editors, *Proc. 35th ACL and 8th EACL*, pages 174–181, Somerset, New Jersey, 1997. ACL.

[8] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proc. 18th International Conference on Computational Linguistics*, 2000.

[9] Daniel Marcu. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448, 2000.

[10] J. R. Martin and P. R. R. White. *The Language of Evaluation: Appraisal in English*. Palgrave, London, 2005. (`http://grammatics.com/appraisal/`).

[11] Christian Matthiessen. *Lexico-grammatical cartography: English systems*. International Language Sciences Publishers, 1995.

[12] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP-2004*, pages 412–418, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[13] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proc. 2nd Int'l Conf. on Knowledge Capture*, 2003.

[14] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. 42nd ACL*, pages 271–278, Barcelona, Spain, July 2004.

[15] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, 2002.

[16] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[17] Y. Qu, J. G. Shanahan, and J. Wiebe, editors. *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. AAAI Press, Stanford University, 2004.

[18] M. Taboada and J. Grieve. Analyzing appraisal automatically. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. AAAI, 2004.

[19] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the ACL (ACL'02)*, pages 417–424, Philadelphia, Pennsylvania, 2002.

[20] Casey Whitelaw, Maria Herke-Couchman, and Jon Patrick. Identifying interpersonal distance using systemic features. In *AAAI Spring Symp. on Exploring Attitude and Affect in Text*, 2004.

[21] J. Wiebe. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. PhD thesis, State University of New York at Buffalo, 1990.

[22] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proc. 19th National Conference on Artificial Intelligence*, 2004.

[23] Ian H. Witten and Frank Eibe. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.