

Using Appraisal Taxonomies for Sentiment Analysis

Casey Whitelaw

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

Navendu Garg and Shlomo Argamon

Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
10 W. 31st Street; Chicago, IL 60616
{gargnav, argamon}@iit.edu

Abstract

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. 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.

1 Introduction

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 (2004)). Second, the “atomic units” of such expressions are not individual words, but rather *appraisal groups*.

2 Appraisal Groups

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* (in English) comprises a *head adjective* with defined attitude type, with an optional preceding list of *appraisal modifiers*, each denoting a transformation of one

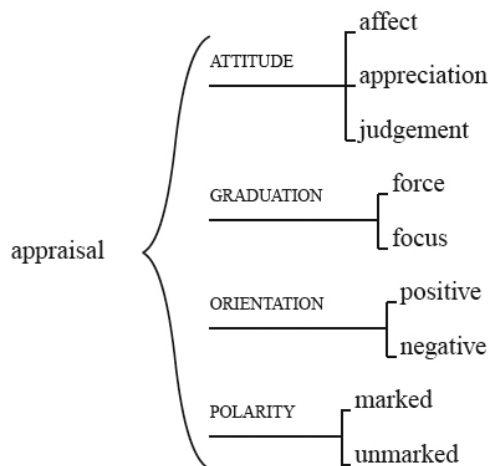


Figure 1: Main attributes of APPRAISAL and their highest-level options.

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.

2.1 Appraisal Group Extraction

Our first goal is to extract appraisal groups, from which we then derive useful feature sets for machine learning. We consider in this paper extraction of a main type of appraisal groups, *adjectival* appraisal groups, which give good classification results despite seemingly low coverage in the corpus. (Fig. 1) shows four main types of attributes are assigned to appraisal groups: Attitude, Orientation, Graduation, and Polarity. The taxonomies for the attributes are adopted from Martin and White’s Ap-

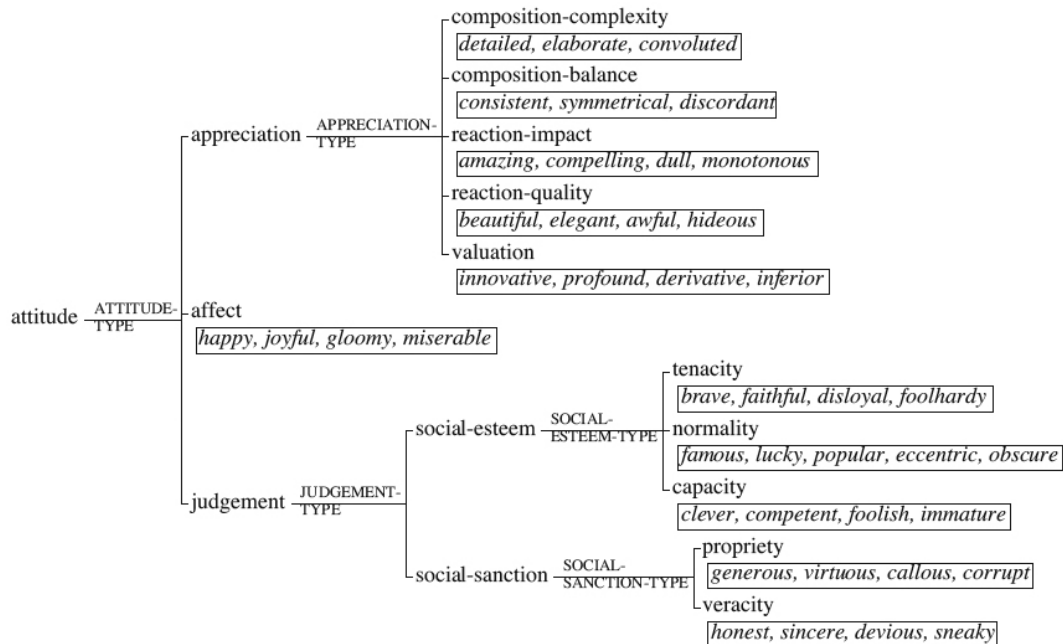


Figure 2: Options in the Attitude taxonomy, with examples of appraisal adjectives from our lexicon.

praisal Theory (2005), developed within the tradition of Systemic Functional Linguistics (Halliday, 1994).

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. Figure 2 gives a more detailed view of the various options in Attitude, together with illustrative 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”.

Figure 3 shows the derivation of the appraisal attributes of “not very happy”.

3 Lexicon

3.1 Lexicon Design

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 adjective¹; for example, the lexical entry for ‘beautiful’ reads:

‘beautiful’	
Attitude:	appreciation/reaction-quality
Orientation:	positive
Force:	neutral
Focus:	neutral
Polarity:	unmarked

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

‘very’	
Force:	increase

¹Force 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.

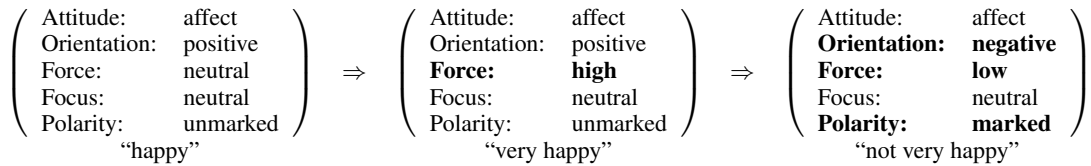


Figure 3: Analysis of appraisal group “not very happy”.

or polarity modification:

‘not’	
Orientation:	negate
Force:	reverse
Polarity:	marked

3.2 Building of the Lexcon

To build the lexicon, we started with the example words and phrases given for various appraisal options in (Martin and White, 2005) and (Matthiessen, 1995) as seed terms, using a semi-automated technique to quickly build a lexicon with decent coverage of adjectival appraisal. Candidate expansions for each seed term were generated from WordNet and from two online thesauri². 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 candidate list. This provides 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.

In total, 1329 terms were produced from 400 seed terms, in around twenty man-hours. Appraisal adjectives in the lexicon cover 29.2% of adjectives in the testbed corpus, comprising 2.7% of words in the corpus³.

²<http://m-w.com> and <http://thesaurus.com>

³Note 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

4 Methodology

4.1 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 text⁴. 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.
- BoW** *Bag of Words* — relative frequencies of all words in the text.

appraisal group analysis to be highly portable to other related tasks.

⁴In preliminary experiments, the use of relative frequencies within each node of the taxonomy, as in (Argamon and Dodick, 2004; Whitelaw et al., 2004), gave inferior results to this simpler procedure.

Feature Set	N_{feat}	Acc. (%)
W:A	1047	77.6
S:A	1355	78.0
S:AO	1278	78.2
G:A	1136	78.2
G:AO	1597	78.6
BoW	48,314	87.0
BoW+G:AO	49,911	90.2
<hr/>		
<i>P&L-04</i>		87.2
<i>M&C-03(TVO)</i>		69.0
<i>M&C-03(best)</i>		86.0

Table 1: Words by Attitude (W:A), Systems by Attitude (S:A), Systems by Attitude & Orientation (S:AO), Appraisal Groups by Attitude (G:A), Appraisal Groups by Attitude & Orientation (G:AO), Bag Of Words (BoW), Pang and Lee’s (2004) (P&L-04), Mullen and Collier (2004) (M&C-03).

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

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

4.2 Corpus

We evaluated the effectiveness of the feature sets for movie review classification, using the publicly available collection of movie reviews constructed by Pang and Lee (2004). This standard testbed consists of 1000 positive and 1000 negative reviews, taken from the IMDb movie review archives⁵.

4.3 Evaluation

Table 1 gives classification results for using WEKA’s (Witten and Eibe, 1999) implementation of the SMO (Platt, 1998) learning algorithm, with default parameters and a linear kernel; accuracy is measured by 10-fold cross-validation. For comparison, we also list previous results from two studies. Directly comparable is Pang and Lee’s (2004) highest 10-fold CV accuracy for this dataset, using ‘subjectivity clustering’ and bag-of-words classification. We also show two results from Mullen and Collier (2004), 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).

The baseline of using just attitude-bearing adjectives (W:A) is reasonably high, at 77.6% accuracy. This bears out our contention that attitude-bearing adjectives specifically are a key feature in the expression of sentiment. Using attitude type and orientation (S:AO) of these terms yields a small improvement in accuracy to 78.2%.

⁵See <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

When using appraisal groups, which include the effect of appraisal modifiers, we see that using both Attitude Type and Orientation(G:AO), we get slightly higher accuracy of 78.6%. The small size of the 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.

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 87% accuracy, competitive with Pang and Lee’s (2004) result on this dataset. More significantly, we improve clearly on that result (attaining 90.2% 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.

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-speech⁶.

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 (Hatzivassiloglou and McKeown, 1997). This is equivalent, in our approach, to using just Orientation values, computing a weighted sum for the whole document.

Semantic orientation has also been useful for classifying more general product reviews (Turney, 2002); 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.

Early attempts at classifying movie reviews used standard bag-of-words techniques with limited success (Pang et al., 2002). The addition of typed features of semantic orientation has been shown to improve results (Mullen and Collier, 2004).

⁶Curiously, ‘and’, ‘also’, and ‘as’ are strong features for positive sentiment. This may indicate that rhetorical structure (Marcu, 2000; Argamon and Dodick, 2004) is also important for understanding sentiment.

There have been some previous attempts at using a more structured linguistic analysis of text for sentiment classification, with mixed results. Mullen and Collier (2004) 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. Previous work on including polarity ("good" vs. "not good") have given inconsistent results—either a slight improvement (Pang et al., 2002) or decrease (Dave and Lawrence, 2003) from bag-of-word baselines; our results show it to help slightly.

Wilson et al. (2004) 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 (2004) 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.

The major challenge facing sentiment classification is the accurate identification of relevant Appraisal Expressions. Without some form of summarization or filtering, performance is limited by the presence of extraneous and potentially misleading appraisal in the document.

References

- S. Argamon and J. T. Dodick. 2004. Conjunction and modal assessment in genre classification. In *AAAI Spring Symp. on Exploring Attitude and Affect in Text*.
- D. Dave and S. Lawrence. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. Twelfth Int'l World Wide Web Conference (WWW2003)*.
- Michael A. K. Halliday. 1994. *Introduction to Functional Grammar*. Edward Arnold, second edition.
- V. Hatzivassiloglou and K. R. McKeown. 1997. 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. ACL.
- Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448.
- J. R. Martin and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave, London. (<http://grammatics.com/appraisal/>).
- Christian Matthiessen. 1995. *Lexico-grammatical cartography: English systems*. International Language Sciences Publishers.
- Tony Mullen and Nigel Collier. 2004. 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. Association for Computational Linguistics.
- Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. 42nd ACL*, pages 271–278, Barcelona, Spain, July.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.
- J. Platt. 1998. 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.
- M. Taboada and J. Grieve. 2004. Analyzing appraisal automatically. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. AAAI.
- Peter D. Turney. 2002. 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.
- Casey Whitelaw, Maria Herke-Couchman, and Jon Patrick. 2004. Identifying interpersonal distance using systemic features. In *AAAI Spring Symp. on Exploring Attitude and Affect in Text*.
- Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proc. 19th National Conference on Artificial Intelligence*.
- Ian H. Witten and Frank Eibe. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.