Analyzing Appraisal Automatically

Maite Taboada and Jack Grieve
Department of Linguistics
Simon Fraser University
Burnaby, B.C., V5A 1S6, Canada
mtaboada@sfu.ca, jwgrieve@sfu.ca

Abstract

We present a method for classifying texts automatically, based on their subjective content. We apply a standard method for calculating semantic orientation (Turney 2002), and expand it by giving more prominence to certain parts of the text, where we believe most subjective content is concentrated. We also apply a linguistic classification of Appraisal and find that it could be helpful in distinguishing different types of subjective texts (e.g., movie reviews from consumer product reviews).

Classifying Sentiment

The task of classifying texts based on their subjective content, or sentiment, is considered to be difficult to implement computationally. There is, however, a growing body of research both in computational and theoretical linguistics that attempts to classify and quantify subjective content. In this paper, we describe our current attempts at designing a system to perform an automatic analysis of sentiment.

We started out by using an existing method for calculating the semantic orientation of adjectives in a text (Turney 2002), but instead of simply averaging the semantic orientation of certain words in the text (in our case, adjectives), we took into account text structure. We also improved on the method by applying Appraisal, a linguistic classification of subjectivity (Martin 2000).

Our system was developed using a corpus of 400 opinion texts, reviews retrieved from the website Epinions.com, divided into 200 classified as "recommended" by the authors (positive), and 200 as "not recommended" (negative). The texts discuss products and services: movies, books, cars, cookware, phones, hotels, music, and computers.

Semantic Orientation for Adjectives

For some time now, researchers have been exploring the subjective content or *semantic orientation* (SO) of words. Hatzivassiloglou and McKeown (1997) proposed a method to quantify the subjectivity inherent in some adjectives, by extracting conjoined adjectives from a corpus: *simple and well-received* are classified in the same set, since they are joined by a coordinating conjunction.

The drawback of classifying and annotating semantic orientation is that either large amounts of data or time-consuming manual coding are needed in order to decide on semantic content. Pang et al. (2002) perform sentiment classification based on machine learning techniques. Turney and Littman (2002) proposed a method in which the Web is used as a corpus. They assumed that a negative word will be found in the neighborhood of other negative words, and conversely for positive words. They used Altavista's NEAR operator, querying the search engine with the word in question and NEAR positive or negative adjectives (ten words in the vicinity of the word). Seven positive (*good, nice, excellent, positive,* etc.) and seven negative adjectives (*bad, nasty, poor, negative,* etc.) were used. They then calculated PMI (Pointwise Mutual Information) for the queries. The result is a positive or negative number, which determines the semantic orientation of the word.

In this first part of the project, we are extracting only adjectives, and calculating their subjective content. Bruce and Wiebe (2000) found that adjectives alone are a good predictor of subjectivity in a sentence. To calculate semantic orientation, we have adopted Turney's method. He uses it (Turney 2002) to calculate SO for adjective+noun, or noun+noun combinations (*cool thing, online experience*).

In our experiment, all reviews were collected from the web site Epinions.com. They are tagged using Brill's tagger (1995) and words with the label JJ are extracted. Some adjectives are discarded: determiner-like adjectives such as *previous* and *other*, and adjectives that had very low hits after a web search (such as misspelled adjectives and novel compounds, e.g., *head-knodding, hypersexual, club-ready*).

The crucial aspect of calculating SO for a text is how to aggregate the values of each adjective. In future work, we plan to parse the text, relying on dependency relations and rhetorical relations (Mann & Thompson 1988) for an accurate portrayal of the relations in the text. We also plan to segment texts that may contain more than one subject topic, and therefore possibly a different SO for each topic or subtopic. For now, we enhanced simple averaging by taking into account text structure, as described in the next section.

Improving Basic SO Classification

Texts are usually structured, at the most basic level, into beginning, middle and end parts. Our hypothesis is that opin-

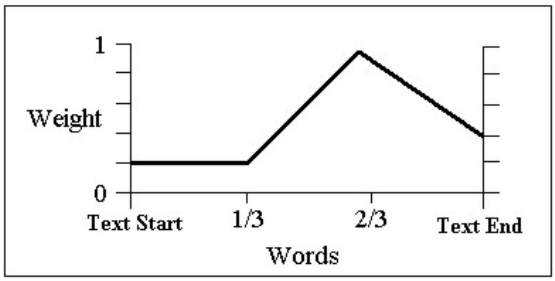

Figure 1: Prominence schema.

	Positive	Negative	Overall
Books	28%	88%	58%
Computers	52%	8%	66%
Hotels	92%	52%	72%
Music	48%	8%	64%
Phones	68%	68%	68%
Movies	32%	88%	6%
Cars	8%	6%	7%
Cookware	96%	28%	62%
All	62%	68%	65%

Table 1: SO accuracy, per review type.

ions expressed in a text will tend to be found in specific parts. Intuitively, those parts should be the middle and the end. Especially in reviews, authors tend to end with a summary statement of their opinion. In order to implement this theory, when analyzing a text, we weight every adjective's SO based on where it occurs in the text.

We experimented with a number of sets of peaks and troughs, defined by four points in the text, and we weighted every word according to this scheme (see Figure 1), so that $word_w = (word_{SO})(weight)$. These weighted SO values were then averaged to determine a text's overall SO. The result was compared to the author's recommendation, a two-point scale (recommended or not recommended). As we expected, the prominence schema that provided the best results was one that disregarded the beginning of the text, increased importance in the second third, and descended towards the final part of the text, as represented in Figure 1.

We also raised the split between negative and positive reviews from 0 to 0.228. Few of the reviews were identified as having negative SO values, and the range of negative values was smaller. The most negative review had a value of 1.5, and the most positive review had an SO of 2. Negative adjectives are used less frequently than their positive counterparts in English (Leech, Rayson, & Wilson 2001).

By averaging weighted SOs, and setting the split between negative and positive reviews at 0.228, we obtained the results in Table 1. The overall accuracy of 65% is a significant improvement over the 51% we obtain if we use simple averages of adjective SO values, and leave the split at zero. There are differences between book, movie and music reviews on the one hand, and phones, cars and cookware on the other. Whereas the system is accurate on the positive reviews for art reviews, it does better on the negative reviews of products. We think this is because product reviews discuss the product in terms of its various components: in car reviews, authors describe brakes, acceleration, safety, appearance. Authors may view a car negatively, even though some of its components receive positive evaluations. The art-related reviews, on the other hand, tend to be more holistic. Although movies can be discussed in terms of their components (score, plot, director, actors), it is possible that the overall subjective evaluation is reflected in each one of those components. It is also important to remember that we are comparing the text content (as represented in its adjectives) to the author's recommendation. The content of the text typically reflects the type of recommendation; however, that is not necessarily the case.

Appraisal

Appraisal is a linguistic theory of subjectivity. The Appraisal system (Martin 2000; 2003), within Systemic Functional Linguistics, is an attempt to model language's ability to express and negotiate opinions and attitudes within text. Martin divides the Appraisal system into three distinct subsystems (see Figure 2): Affect, Judgement, and Appreciation, which model the ability to express emotional, moral, and aesthetic opinions respectively. The speaker's use of a specific set of words associated with each of these subsystems will be applied in the current study to locate and evaluate Appraisal in text.

In addition to these three sub-systems, Martin argues that two other systems play a crucial role in the expression of opinion. The Engagement system is the set of linguistic options that allow the individual to convey the degree of his or her commitment to the opinion being presented. And the Amplification System is responsible for a speaker's ability to intensify or weaken the strength of the opinions they express. Figure 2 summarizes the Appraisal systems. The curly bracket indicates simultaneous choice, and the square bracket exclusive choice. The examples represent a typical use of the adjective in evaluative text.

Appraisal is an addition to our semantic orientation based review classification system, not a substitution. We believe that Appraisal will help us categorize the opinions contained in a text, and whether they refer to objects, emotions or behaviors. Additionally, using Amplification and Engagement, we will be able to quantify the writer's commitment to the opinion, and how focused that opinion is. A text will be assigned an SO value, and one or more Appraisal values. The SO value could be compared to Amplitude in Appraisal terms: to what extent the text is positive or negative.

Analyzing Appraisal

Once we determine whether a review is of negative or positive orientation, the next step is to determine the degree to which a review expresses Affect, Judgement and Appreciation. For example, a review of an anti-depressant drug would

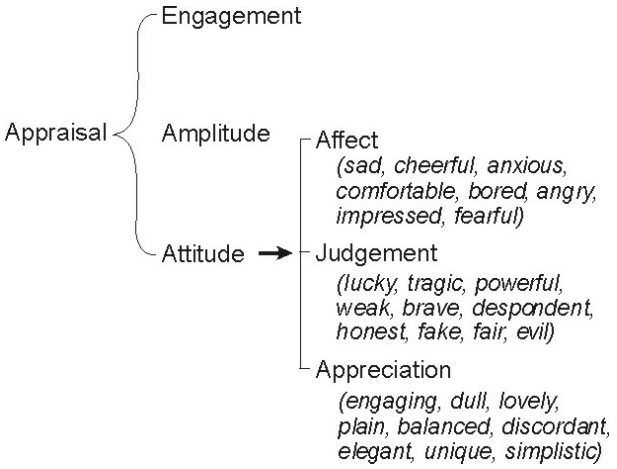

Figure 2: Appraisal systems.

mostly contain Affect, while a review of a restaurant's service would consist of Judgement, and a literary review, of Appreciation. A combination could be used too: a concert reviewer might consider the quality of the music (Appreciation) but also comment on the showmanship of the musicians (Judgement).

We consider the potential of one such method for Appraisal-based review classification. Like Turney's method for measuring a review's semantic orientation, this method is based on adjective frequency. Except that here, it is assumed that a review's degree of Affect, Judgement and Appreciation can be determined by counting adjectives used to express each type. If every adjective was used only to express one of these three basic types of evaluation, then we would simply need to compile three lists of adjectives: if a document was found to contain four Affect adjectives, five Judgement adjectives and one Appreciation adjective, it would be deemed to be 40% Affect, 50% Judgement and 10% Appreciation (an evaluation, perhaps, of a preacher or of a politician). Of course, this is not the case: adjectives have the potential to express Affect, Judgement and Appreciation depending on the context in which they are used. We must therefore find some way to determine an adjective's overall *evaluative potential*—the probability that an adjective will be used in evaluative discourse to express Affect, Judgement or Appreciation.

In Table 2 we present ten examples of adjectives whose evaluative potential has been estimated based on the adjective's definition, its most frequent collocations in the British National Corpus, and our own intuitions about its usage in evaluative texts. *Afraid*, for example, when used to express opinion, tends to be used as an Affect adjective, as in *I left the movie theatre so afraid that I ran all the way home*. But it can also be used to express Judgement: *The police department is useless: they are too afraid of criminals to arrest them*, or, on rare occasions, to express Appreciation: *The man in the painting looks afraid*. *Cute*, on the other hand is generally used as a Judgement adjective (opinion about another person's looks and behavior). But one could also say that *The dress is cute* (Appreciation) or even *The dress makes me feel cute* (Affect).

Adjective	Affect	Judgement	Appreciation
Afraid	0.6	0.3	0.1
Aware	0.5	0.4	0.1
Cute	0.1	0.6	0.3
Great	0.1	0.2	0.7
Happy	0.6	0.3	0.1
Intelligent	0.2	0.7	0.1
Little	0.1	0.2	0.7
Quick	0.1	0.8	0.1
Red	0.1	0.2	0.7
Weak	0.3	0.5	0.2

Table 2: Assigned Appraisal values for ten adjectives.

In total, such estimations were made for fifty frequent adjectives drawn from our semantic orientation database. We then tested various methods for determining an adjective's evaluative potential automatically—all of which relied on web-based mutual information calculations like those developed by Turney to estimate an adjective's semantic orientation—to see which would compute evaluation potential which were closest to our own estimations. The best method turned out to be surprisingly simple. In order to calculate an adjective's evaluative potential first the mutual information between the adjective and the pronoun-copula pairs *I was*, (Affect); *he was*, (Judgement); and *it was* (Appreciation) were calculated using formula (1) and the search engine at AltaVista.com, where PRO stands for one of the pronouns: *I/he/it*. The adjective's potential to express Affect, Judgement and Appreciation was calculated using formulas (2)-(4).

1. MI(*PRO was*, **A**) = $\log_2$(hits(*PRO was* **A**) / hits(*PRO was*) hits(**A**))
2. Affect Potential = MI(*I was*, **A**) / (MI(*I was*, **A**) + MI(*he was*, **A**) + MI(*it was*, **A**))
3. Judgement Potential = MI(*He was*, **A**) / (MI(*I was*, **A**) + MI(*he was*, **A**) + MI(*it was*, **A**))
4. Appreciation Potential = MI(*It was*, **A**) / (MI(*I was*, **A**) + MI(*he was*, **A**) + MI(*it was*, **A**))

The results of these calculations for the ten adjectives, whose Appraisal potential was estimated above, are presented in Table 3. While this method easily outperformed all other methods, its results were by no means perfect. In particular, the most prevalent problem involved common but misleading collocations: the evaluative potential of a clear Appreciation adjective such as *little*, for instance, is misidentified as being an Affect adjective because *when I was little* is such a common phrase. But since, overall, this method was found to yield the most consistent and accurate results, we are adopting this method over the coding based on our own intuitions, since it can be performed automatically for a large number of adjectives.

Before we test to see if these values are useful in determining a review's degree of Affect, Judgement and Appreci-

Adjective	Affect	Judgement	Appreciation
Afraid	0.66	0.34	0.00
Aware	0.44	0.54	0.02
Cute	0.12	0.44	0.44
Great	0.01	0.11	0.88
Happy	0.67	0.32	0.01
Intelligent	0.16	0.77	0.07
Little	0.71	0.18	0.11
Quick	0.15	0.72	0.13
Red	0.14	0.25	0.61
Weak	0.39	0.51	0.10

Table 3: Appraisal values from corpus.

Subject	Affect	Judgement	Appreciation
Books	23	27	50
Computers	20	24	56
Hotels	21	26	53
Music	22	28	50
Phones	17	22	61
Movies	23	26	51
Cars	20	23	57
Cookware	19	24	57

Table 4: Appraisal values per review type.

ation, it is worth first briefly considering the relevance of this method for estimating an adjective's evaluative potential to Martin's theory of Appraisal. In particular, we can now offer a grammatical justification for why Martin's three categories of Appraisal seem to be sufficient. Since adjectives modify nouns, and since there are three basic types of nouns—objects, people, and one's self, as confirmed by the range of pronouns by which they may be replaced—it should not be surprising that when an opinion is expressed about some thing, that the opinion itself can take one of three forms: it can comment on a thing (Appreciation), a person (Judgement), or one's self (Affect). Martin (2003) does point out that the three categories can be expressed as *I feel x* (Affect), *It was x of him/her to do that* (Judgement), and *I consider it x* (Appreciation). We have simplified and generalized the frames to make them refer to the self, persons and things.

We now turn to the evaluation of our Appraisal-based method classifying reviews. In Table 4, we provide the average Appraisal values for each subsection of our review corpus. The figures indicate the percentage content of each type of Appraisal. We see, as with the SO results, that the reviews seem to fall into different classes. Cars and consumer products (computers, phones and cookware) have higher Appreciation values. These refer to quality, cost, and manufacturing. There is a class that also has high Judgement values: books, music, and movies, the art-related items. Finally, hotels have both high Judgement and Appreciation.

We could use Appraisal categorization to determine whether a review is about a consumer product or not (high Appreciation). We could also use it to extract the most relevant reviews: when looking for hotels, a consumer might be interested in the elegance of the rooms (Appreciation), or the service (Judgement). A system could extract hotel reviews with high values of one or the other, depending on the user's preferences. This is a tentative proposal; the corpus is too small to draw further conclusions.

Conclusions

Our first approach to classifying semantic orientation for adjectives yielded encouraging results. We expanded on a basic method for extracting semantic orientation by taking into account the position for each of the adjectives in the text. We found that the performance of our system varies depending on the type of review under consideration. We extracted Appraisal values for each of the reviews, also revealing different characteristics according to review type.

Future work will deal with negation in a principled way (we don't think that *not excellent* should have the reverse value of *excellent*), calculating values for words other than adjectives, and using collocations, rather than single words.

Acknowledgments

This project was supported by a grant to the first author from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Katia Dilkina for implementing part of the semantic orientation procedure, and Dennis Storoshenko for assisting with data collection.

References

Brill, E. 1995. Transformation-based error-driven learning and Natural Language Processing. *Computational Linguistics* 21(4):543–565.

Bruce, R., and Wiebe, J. 2000. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering* 5(2):187–205.

Hatzivassiloglou, V., and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of 35th ACL.*, 174–181.

Leech, G.; Rayson, P.; and Wilson, A. 2001. *Word Frequencies in Written and Spoken English: Based on the National Corpus*. London: Longman.

Mann, W., and Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3):243–281.

Martin, J. 2000. Beyond Exchange: Appraisal systems in English. In Hunston, S., and Thompson, G., eds., *Evaluation in Text*. Oxford: Oxford University Press. 142–175.

Martin, J. 2003. Introduction, Special issue on Appraisal. *Text* 23(2):171–181.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using Machine Learning techniques. In *Proc. Conf. on Empirical Methods in NLP*, 79–86.

Turney, P., and Littman, M. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th ACL.*, 417–424.