Appraisal Extraction for News Opinion Analysis at NTCIR-6

Kenneth Bloom Sterling Stein Shlomo Argamon
Department of Computer Science
Illinois Institute of Technology
10 W 31st St. Chicago, IL 60616
kbloom1@iit.edu stein@ir.iit.edu argamon@iit.edu

Abstract

We describe a system which uses lexical shallow parsing to find adjectival "appraisal groups" in sentences, which convey a positive or negative appraisal of an item. We used a simple heuristic to detect opinion holders, determining whether a person was being quoted in a specific sentence or not, and if so, who. We also explored the the use of unsupervised learners and voting to increase our coverage.
Keywords: *Appraisal theory, opinion extraction.*

1 Introduction

Our entry to the NTCIR opinion track is based on our appraisal extraction system which applies the attitude system from Martin and White's [4] Appraisal Theory. An *appraisal expression* is an elementary unit of text by which an opinion holder (the *source*) expresses an opinion (the *attitude*) about a *target*. In an appraisal expression, the three functions of source, attitude, and target may not be found contiguously in the text (instead being connected syntactically), and some functions (like source or target) may not be explicit, left by the speaker to be inferred from context.

Appraisal Theory [4] is a grammatical theory dealing with how opinion is represented in text. The attitude system classifies evaluative language into three general types of opinions: *affect* (an internal emotional state), *appreciation* (of intrinsic qualities of an object), or *judgment* (concerning the way people behave). English grammar imposes different constraints on how these three types of appraisal can be expressed. One cannot, for example, talk about "an evil towel" very easily because "evil" is a type of judgment, but a towel is an object that does not have behaviors (unless anthropomorphized). Similarly, to say "Alice is nice" is very different from saying "Alice is happy" because "nice" and "happy" are two different types of appraisal. "Nice" is used to make a judgment about Alice's typical behavior, whereas "happy" is affect, describing an Alice's emotions, and Alice functions as the emoter.

Our version of opinion extraction is based on Appraisal Theory, extracting the parts of appraisal expressions which are relevant to the NTCIR opinion task, in this case the source and the attitude. Our approach to this task involved building a general lexicon of words that can be used to express attitudes, and shallow parsing to find whole phrases (which may carry different orientation than the single words listed in the lexicon). Thus far, we have only attacked a simplified version of the problem, using the system to detect adjectival and adverbial attitude groups. We are beginning work on detecting nominal and verbal attitude groups, but previous work has shown us that even without these there is still a lot of information to be gleaned from adjectival and adverbial attitude groups.

We have applied this system to the problems of movie review classification [6], and to a form of opinion mining intended to aggregate public opinion concerning specific parts of products [1, 2].

It seems that the task's notion of opinion extraction differs somewhat from appraisal extraction. While it is easy to see that everything described by attitude could be construed as opinion, the opposite is not necessarily true. For example, consider the sentence

> Isn't it time Switzerland reconsidered its tenacious neutrality and joined the [European] union?

This sentence was considered to contain opinion by two of the three NTCIR raters. Although "tenacious neutrality" is an example of explicit appraisal, the sentence would still be opinionated even if the word "tenacious" was removed, thereby eliminating the appraisal expression from the sentence.

Indeed, we found it difficult to determine from the provided sample data (or from the task description) a general definition of what NTCIR raters considered to be opinion and what they did not. Presumably the goal should be to detect expressions that are construed in the text as opinions rather than as facts. But without a clear, theoretically motivated delineation of this con-

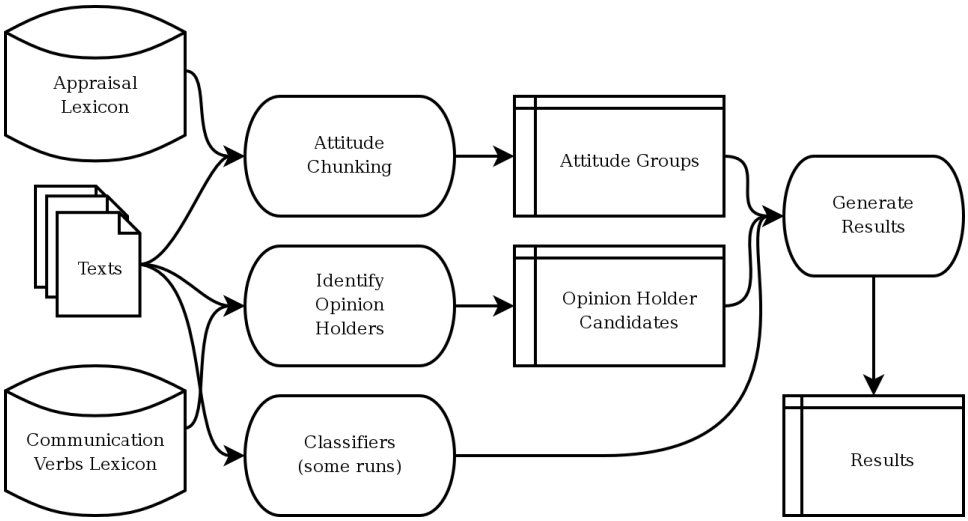

Figure 1. Appraisal Extraction architecture

cept, reasonable people may often disagree on whether something is construed as an opinion or not. This is borne out by our evaluation of the agreement percentages of the raters for subjectivity and polarity classification. We found average pairwise interrater agreement (kappa) for subjectivity[1] to be 0.23, and for polarity[2] to be 0.27, both of which indicate only low levels of agreement beyond what chance would predict. These results indicate that more work may be needed on developing better formulations of the notion of "opinion" that will enable more consistent human annotation.

In this paper, we describe how we implemented our opinion extractor. In section 2, we discuss our lexicon of appraisal words. In section 3, we explain how these appraisal words are used to find "appraisal groups". Section 4, details our method for detecting opinion holders on a sentence-wide scale. In section 5, we discuss our use of several machine learning classifiers to go beyond the static lexicon. In section 6, we explain our various output runs. In section 7, we discuss our submitted results and several other results.

2 Lexicon

To identify the various parts of the appraisal expressions, we built a generic lexicon of words and phrases used to express appraisal attitudes. We used the lexicon developed for our previous work [1, 6] on appraisal.[3]

The original lexicon comprised entries for appraisal *heads*, such as 'good', 'bad, and 'ugly', and also *modifiers*, such as 'very', 'somewhat', and 'truly'. The lexicon lists values for the attributes attitude type, force, focus, orientation, and polarity for head words, and lists modification operations on the same attributes for modifiers. The only attribute that we were concerned with for NTCIR was orientation, which indicates whether an appraisal group is positive or negative. Most words in the lexicon are either positive or negative – only two out of about two thousand words are considered neutral.

The polarity attribute indicates whether an appraisal expression's orientation is flipped by a modifier like "not."[4] We were not concerned with value of the polarity attribute as listed in the lexicon, since changes in polarity always change the orientation as well (as in the example in Figure 2). We discuss in section 5 a classifier intended to determine whether there were polarity markers that the system could not capture lexically through chunking.

We added to that lexicon attribute values for adjectives culled from the full NTCIR corpus to increase the coverage of the lexicon. In order to capture at least a little bit of the appraisal that can be conveyed non-adjectivally, we also incorporated Levin's [3] lists of admire-type verbs and judgement verbs.

We constructed a second lexicon of communication verbs which we used to locate opinion holders in the

[1] Cohen's kappa. We note that the average interrater agreement percentage was 72%, but the distribution is highly skewed.The stringent gold standard is 5% opinionated, and the lenient gold standard is 25% opinionated, so there is high agreement on a large number of non-opinionated cases, but less agreement on opinionated cases.

[2] Fleiss' kappa computed on the entire 4x4 polarity confusion matrix. We note that the average agreement percentage was 69%, but the confusion matrix is again highly skewed.

[3] The lexicon used for this work is available at `http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz`

[4] Our usage of the term polarity in this way is for consistency with the terminology used in Systemic Functional Linguistics. In section 5, we use the term to discuss a classifier intended to detect this kind of negation, but elsewhere in this paper we use it more or less interchangeably with orientation (particularly when discussing our results) for consistency with the NTCIR task terminology.

text. These were also constructed based on Levin's [3] lists, specifically her lists of verbs similar to 'characterize', 'declare', 'conjecture', 'admire', 'judgement', 'assess', 'say', 'complain' and 'advise'.

We matched the communication verbs in the corpus by using a Porter stemmer to stem the lexicon and the corpus so that we could account for all of the verb forms by matching the Porter stem. To match the same verbs as appraisal, we simply listed all of their verb forms in the appraisal lexicon, so as to avoid modifying software that we had already written to operate without using a stemmer. Thus, for the verb "chide", we listed the verb forms "chide","chided","chides", and "chiding" in the appraisal lexicon, and just the base form "chide" in the communication verbs lexicon. As a special exception, we added the past tense "said" to the communication lexicon because it is clearly a very common verb, but the Porter stem of "said" does not match the Porter stem of "say".

3 Chunking

Chunking is the process of finding attitude groups, and is performed according to a technique discussed in our previous work [6]. An *adjectival attitude group* (in English) comprises a *head adjective* with defined values for its attributes from the lexicon, with an optional preceding list of *appraisal modifiers*, each denoting a transformation of one or more appraisal attributes of the head. For example, 'not very nice', has head 'nice' and modifiers 'not' and 'very'. We take advantage of typical English word-ordering and use all pre-modifiers, allowing for intervening articles. This allows groups such as 'really *a* very beautiful...', where 'really' is taken to modify 'beautiful'. Transformations to appraisal attributes are applied with right associativity, so the phrase 'not very good' is transformed by starting with the word 'good', then applying the modifier 'very' then applying 'not'. An example of this process is shown in Figure 2.

4 Source Detection

Our approach to extracting opinion holders focused on determining who was the "primary opinion holder" in any given sentence by looking at changes in who the news article was quoting at different times. We create a list of all potential opinion holders for all sentences in the corpus, by finding the names of those who are quoted at the beginning or end of each quote, and then applying them to the other sentences in the quote. We find these potential opinion holders by using the lexicon of communication verbs described in Section 2. The opinion holder, therefore, was either the subject of these verbs (when active) or the agent of these verbs (when passive, for example "said by John"). Using a dependency parse of the corpus, we identified the subject of these verbs by following the appropriate syntactic links, and we used shallow parsing to identify the whole name of the subject.

Once these opinion holders are identified for the sentences in which they are explicitly mentioned, we identify the beginnings and endings of quotes by balancing the quotation marks in the document. The system tracks the total number of quotation marks encountered in a document when reaches the end of the each sentence. If this number is odd then the sentence is considered to be within a quote (or the beginning of a quote). If this number is even, but the sentence contains a quotation mark in it, then the sentence is considered to be the end of a quote. The system does not look at all at curly quotes already in the document, which would differentiate between opening quotation marks (“) and closing quotation marks (”), since curly quotes were not used by all newspapers in the corpus.

This technique works for the NTCIR corpus because the corpus consists of news articles edited by professional editors who enforce this convention. There was only one article in the corpus where the editor made a mistake, and two articles that broke this convention by having multi-paragraph quotes where every paragraph starts with a quotation mark, but only the last paragraph ends with one. With other kinds of data, such as blogs or user contributed product reviews, we would need to consider other methods of quote detection.

In the sample data, we found two interviews which needed to be treated specially. In an interview, a new speaker's response is signaled by starting the sentence with the speaker's name, followed by a colon. The speaker may say several sentences before the next speaker starts, but typically every opinion expressed in these sentences is the opinion of the speaker of these sentences. We therefore developed a technique for detecting interviews and sharing opinion holders between sentences. The system detected three interviews that used this format. There was at least one interview in the corpus that had originally used different text styles to denote the interviewer and the interviewee, which our system was unable to detect as this information was not preserved in the corpus.

We detect an interview by looking for a colon as the second, third or fourth token in a sentence. If there are more than 5 examples of this, then the document is considered an interview. In the examined documents, colons were uncommon in most other documents. Usually the colons were early in the sentence because after the first use of the full name, it is abbreviated.

We identify the speakers in an interview by locating sentences where first colon is the second, third, or fourth token. All of the tokens before that colon are kept as the source. In one interview where the inter-

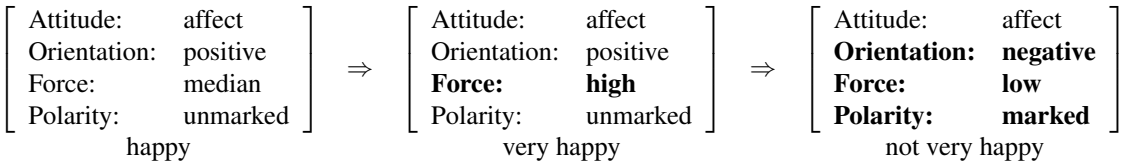

Figure 2. An example of chunking the phrase "not very happy."

viewer and interviewee were denoted by a single letter ("C:" and "M:"), only that single letter is identified as the source.

Once we have the speakers who are explicitly identified by the first sentence of the quote in an interview, they are assigned to all subsequent sentences until the next sentence that begins with a new speaker. Interviews are also subject to the quote extraction method used on all articles.

The system makes no attempt to resolve coreferences.

5 Classifiers

In an attempt to capture lexical regularities not accounted for by our hand-built appraisal system, we experimented with several automatic bootstrapping classifiers trained from the appraisal system to try to achieve better results than the hand built system. The general outline for training the classifiers is as follows:

1. Select appraisal expressions or sentences based on a specific to the classifier being trained. Generate features from these expressions and train a Support Vector Machine from only these instances.

2. Classify all appraisal expressions or sentences based on the classifier.

3. Train a new classifier on on the 10% of the positive reviews with the highest confidence, and the 10% of the negative reviews with the highest confidence.

4. Repeat steps 2 and 3.

All three of these classifiers needed to be trained on a corpus other than the NTCIR corpus, because they depend on features of the appraisal target, or they require overall review classifications in order to bootstrap. Thus, we trained the classifiers on a standard corpus of 2000 IMDB movie reviews developed by Pang and Lee [5], and used the last trained model to classify the NTCIR corpus.

The only classifier that we used to submit any runs was the subjectivity classifier. (We used it in one of the two submitted runs.) The others are described here for completeness — they performed worse over the sample data when they were used, so they were not used to generate any submitted runs.

5.1 Polarity

The polarity classifier is designed to determine whether the orientation and polarity of an appraisal group are correct, or whether their context determines that they should be reversed.

It depends on the technique used in our previous work [1] to identify appraisal targets, and it trains using appraisal expressions whose target is "this-movie". If the appraisal group's orientation disagrees with the overall review classification (for example a negative appraisal group referring to the movie as a whole, while the review itself was positive), then we use this as an example of where the polarity needs to be flipped. In a case where they agree (positive appraisal group, positive review), we use this as an example of where the polarity does not need to be flipped.

Because the NTCIR corpus does not have overall review orientations marked (nor is there an easy analog in the domain of news articles), the classifier was trained on movie reviews corpus used by Pang and Lee [5], which categorizes reviews as positive or negative based on the number of stars the reviewer gave the movie.

The features used in the vectors are 1-,2-, and 3-grams made up of the 5 words immediately preceding the appraisal group.

5.2 Subjectivity and Orientation

The subjectivity classifier is used to classify sentences to determine whether they are subjective or objective.

We had intended that the features used to classify sentences would be a simple bag-of-words feature set, omitting function words and words which appeared in the appraisal lexicon (so that known appraisal words would not become the top features affecting subjectivity, thus causing the classifier to exactly match the appraisal extraction). Due to a coding oversight, we wound up using a feature set that was a holdover from the polarity classifier: 1-,2-, and 3-grams made up of the 5 words at the end of the sentence, skipping over

known appraisal words. We only discovered this bug after the data submission deadline, so after receiving results, we generated a fixed version of the affected run, and we report its results in this paper.

The initial set of positive examples were sentences from the IMDB corpus where an appraisal expression was found, and the computer was able to identify an appraisal target. In normal operation, the computer cannot always identify an appraisal target. This classifier assumes that when the computer could not identify an appraisal target that the attitude group was a less reliable example of real attitude, and so those appraisal expressions are not used as training examples.

The orientation classifier learns to classify sentences as positive or negative. The initial set of positive examples was sentences which contain positive appraisal expressions where a target was found (but containing no negative appraisal expressions), and the initial set of negative examples was sentences containing negative appraisal expressions where a target was found (but containing no positive appraisal expressions). The same feature set was used for orientation as for subjectivity.

6 Output Runs

We submitted two runs for competitive evaluation.

IIT-1 did not use any learned classifiers. Subjectivity and orientation were determined directly from the lexicon and chunking, and no classifier was used to determine whether orientations should be reversed based on context. If a sentence had multiple appraisal expressions, the majority orientation was used. Sentences with equal numbers of positively and negatively oriented appraisal groups were marked neutral.

We then ran two runs using the subjectivity classifier, one trained for five iterations, and one trained for one iteration. In these runs the subjectivity of each sentence was determined by the classifier, and the orientation of each sentence was determined by the attitude groups found in the sentence, using the majority as in run IIT-1. If a sentence had no attitude groups in it, it was marked as neutral.

The other submitted run, **IIT-2**, was created by voting, taking the majority of run IIT-1, and the two subjectivity runs on all decisions regarding subjectivity and orientation. Since opinion holders were the same for all runs, the opinion holders for subjective sentences were taken from whichever of the three runs had marked the sentence as subjective.

We ran various other runs on our with the various other combinations of the classifiers, and compared their results against the ground truth we were given for the sample data when deciding which runs to submit.

Because the IIT-2 run was affected by the bug in the subjectivity classifier mentioned above, we also present here the results of an identical run done with the corrected feature set for the subjectivity classifier. This run is named **IIT-2-Fixed**

Table 1. NTCIR Results for Strict evaluation, requiring all three raters to agree. Italicized values are results that we computed ourselves using NTCIR's scripts.

		IIT-1		IIT-2	
Task Name	Metric	Score	Rank	Score	Rank
Opinion	Precision	0.07015	3 of 9	0.05597	9 of 9
Opinion	Recall	0.57845	7 of 9	0.84000	3 of 9
Opinion	F measure	0.12513	5 of 9	0.10495	9 of 9
Polarity	Precision	0.027	2 of 7	0.016	4 of 7
Polarity	Recall	0.322	2 of 7	0.359	1 of 7
Polarity	F Measure	0.049	2 of 7	0.031	4 of 7
OpHolders	Precision	0.57267	2 of 6	*0.51234*	
OpHolders	Recall	0.46136	2 of 6	*0.58588*	
OpHolders	F Measure	0.51102	2 of 6	*0.54664*	

7 Results

We report in tables 1 and 2 a few selected measures from NTCIR's published results, as well as our ranking relative to all submitted runs. In table 3, we report our own computed results for the IIT-2-Fixed run.

We have selected several key measures to report:

Opinion reports accuracy at identifying opinionated sentences.

Polarity reports accuracy at identifying the correct orientation of a sentence. We list the precision, recall, and F-measure for identifying all polarities, over all sentences, not just sentences that were opinionated in the NTCIR gold standard.

OpHolders reports opinion holders correct relative to all sentences in the corpus, for recall. For precision, this reports opinion holders correct relative to sentences that the system decided were opinionated, even if the NTCIR raters decided they were not opinionated.

We computed opinion holders results for the IIT-2 and IIT-2-Fixed runs by ourselves, using the evaluation scripts provided by NTCIR. Since NTCIR did not compute these numbers themselves, they are indicated on the results tables by italics. NTCIR's evaluation scripts ask a human evaluator to evaluate the opinion holder string matches, such as determining whether "President Clinton" matches "Clinton", or whether "President" matches "the President". We simply answered 'no' to all new string matches when evaluating the IIT-2 and IIT-2-Fixed runs.

7.1 Analysis

In the IIT-1 run, we note a high rank for precision in determining whether a sentence is opinion-

Table 2. NTCIR Results for lenient evaluation, requiring two of the three raters to agree. Italicized values are results that we computed ourselves using NTCIR's scripts.

		IIT-1		IIT-2	
Task Name	Metric	Score	Rank	Score	Rank
Opinion	Precision	0.32491	1 of 9	0.25902	8 of 9
Opinion	Recall	0.58818	7 of 9	0.85375	3 of 9
Opinion	F measure	0.41859	5 of 9	0.39745	8 of 9
Polarity	Precision	0.120	2 of 7	0.086	5 of 7
Polarity	Recall	0.287	2 of 7	0.376	1 of 7
Polarity	F Measure	0.169	2 of 7	0.140	3 of 7
OpHolders	Precision	0.48270	2 of 6	*0.44297*	
OpHolders	Recall	0.40874	1 of 6	*0.53591*	
OpHolders	F Measure	0.44265	2 of 6	*0.48502*	

Table 3. IIT-2-Fixed run. The rank for this run is computed out of all 9 submitted runs, plus this additional run for a total of 10 runs.

		Strict		*Lenient*	
Task Name	*Metric*	*Score*	*Rank*	*Score*	*Rank*
Opinion	*Precision*	*0.076*	*2 of 10*	*0.331*	*1 of 10*
Opinion	*Recall*	*0.496*	*10 of 10*	*0.478*	*10 of 10*
Opinion	*F measure*	*0.131*	*3 of 10*	*0.391*	*9 of 10*
Polarity	*Precision*	*0.022*	*3 of 8*	*0.108*	*3 of 8*
Polarity	*Recall*	*0.210*	*5 of 8*	*0.205*	*6 of 8*
Polarity	*F Measure*	*0.040*	*3 of 8*	*0.141*	*3 of 8*
OpHolders	*Precision*	*0.486*	*4 of 7*	*0.451*	*4 of 7*
OpHolders	*Recall*	*0.336*	*5 of 7*	*0.305*	*5 of 7*
OpHolders	*F Measure*	*0.398*	*4 of 7*	*0.364*	*4 of 7*

ated, while nevertheless achieving low recall. Analysis of the errors validates our intuition that appraisal detection differs somewhat from opinion detection. In other cases, our current restricted implementation of appraisal (with mainly adjectives) hurt recall as well. Nonetheless, since adjectival appraisal is a type opinion, we achieved relatively high precision on what our system extracted. Errors affecting precision were mostly the result of a highly lexical approach to extracting appraisal, which cannot detect when a lexicon word is being used in a non-appraisal context, as is possible with many important appraisal words.

We achieved much higher accuracy at determining the orientation of each sentence. The largest problem our system encountered here is the dearth of neutral words in the appraisal lexicon. Since our system (with substantially the same lexicon) has performed well on a corpus of movie reviews, this suggests that people are much more likely to write positively or negatively opinionated text in a review than in a news article. It may be that appraisal analysis is more useful for reviews than for news articles.

The IIT-2 run boasts high recall and low precision, suggesting that the two automatically trained classifiers identified a lot of subjective text, but were overly zealous in selecting opinionated sentences, but had little correlation to the actual opinionated text. This is not surprising, given the buggy feature set that it used to classify sentences. The IIT-2-Fixed run has higher precision than IIT-1, but lower recall, suggesting that that the three detectors which voted were highly uncorrelated, but when they agreed it was more likely on something that was actually opinionated.

When extracting opinion holders, false positives are usually due to cues listed in the communication verbs lexicon which nevertheless do not signal an opinion holder in context. As we did not tune the communications lexicon for precision, it may be that there are verbs in the communications lexicon that always cause false positives.

Since all of our runs, IIT-1, IIT-2, and IIT-2-Fixed used the same method for detecting potential opinion holders, but opinion holders were only listed when the sentence was listed as opinionated, the differences in performance at extracting opinion holders between the various runs are due entirely to differences between the three runs in detecting opinionated sentences.

8 Conclusions

Despite the slight mismatch between our system's goals and the NTCIR task's, it achieved comparatively high precision in determining opinionated sentences. Recall was adversely affected, however, by the fact that the system currently relies on a fairly small lexicon of adjectives and verbs. This success suggests that this kind of lexicon-based approach may have advantages in precision complementary to those of more fully-automated approaches in recall; hybrid approaches should therefore be explored as well. Regarding determination of opinion polarity (orientation), our use of syntactic relations to extend bag-of-word approaches leads to quite good results, despite our lack of a good model of neutral polarity. We note that our straightforward heuristic for determining opinion holders did quite well overall.

It is possible that the engagement system of Appraisal Theory may be closer to what the NTCIR task considered to be opinion. The engagement system concerns how writers position their beliefs with respect to opinions that they state, and how they position those opinions with respect to alternatives. Included in this system are distinctions regarding how writers express facts, and whom they attribute the facts to. We intend to implement a system to analyze engagement, and expect that this that this may improve results on a similar task.

Division of Labor The core of the appraisal extraction system is described in our previous work [1, 6]. For NTCIR, Argamon supervised and directed the project. The bulk of

the development and evaluation work was done by Bloom. Stein developed the first version of the source detector.

References

[1] K. Bloom, N. Garg, and S. Argamon. Extracting appraisal expressions. *Proceedings of Human Language Technologies/North American Association of Computational Linguists*, 2007.

[2] N. Garg, K. Bloom, and S. Argamon. Appraisal navigator. *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 727, 2006.

[3] B. Levin. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London, 1993.

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

[5] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, US, 2002. Association for Computational Linguistics.

[6] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal taxonomies for sentiment analysis. In *ACM SIGIR Conference on Information and Knowledge Management*, 2005.