A SYSTEM FOR AFFECTIVE RATING OF TEXTS

Stephen D. Durbin, J. Neal Richter, and Doug Warner

RightNow Technologies, Bozeman, MT

1. **Introduction**

In pursuit of automated text understanding, two broad types of approach can be distinguished: analytic methods (e.g. named entity extraction) that provide specific items of information, and synthetic methods (e.g. topic identification) that provide a global characterization. Recent interest in identifying overall affect or sentiment in text falls into the second category. Judging from the limited results reported so far, it appears to be a more challenging problem than topic identification. This is presumably because topic, to first approximation, can reasonably be represented by the straightforward accumulation of word content, whereas tone or affect -- like meaning itself -- depends on relationships of words with each other and with referents external to the text.

Previous work in this area has followed a variety of approaches. One major lineage in the field of psychology stems from the General Inquirer (Stone *et al* 1966), based on recognition of keywords and context-sensitive disambiguation rules, all manually compiled. In part because of the burden of creating the rules in particular, later research tends to use simpler word counting techniques, for example based on the LIWC dictionary (see Pennebaker, Mehl and Niederhoffer 2003 for a review). A similar multi-dimensional view of affect, also based on an assortment of word lists, incorporated fuzzy logic to reflect the multiple valences of many words (Subasic and Huettner 2000).

More recently, artificial intelligence techniques have been applied to the problem, usually simplifying it to determining whether a particular text has predominantly positive or negative affect. This binary text classification has been approached as a supervised learning problem, using methods such as decision tree (Spertus 1997), naive Bayes, maximum entropy, and support vector machine methods (Pang, Lee and Vaithyanathan 2002). To avoid the necessity of a labeled training set, Turney (2002) introduced the idea of calculating the "semantic orientation" of terms or phrases via their difference in mutual information with the words "excellent" and "poor;" the semantic orientation of a text is then the average of that of the terms it contains.

With the exception of the General Inquirer rules, the methods mentioned above ignore syntax, treating a text as a bag of words. Thus, differences in word connotation depending on local context (e.g. "screw") or linguistic community ("bad"), and syntactic effects, especially negation ("not ... happy") are not well handled. A first step beyond the bag of words view is to associate subjects and affective judgments according to their proximity (Tong 2001; Morinaga et al 2002). One would hope that use of more sophisticated natural language processing tools would enhance performance. However, attempts to introduce even simple NLP techniques was reported to provide little increase in accuracy (Kushal, Lawrence, and Pennock 2003). A different line has been taken in recent work that aims to incorporate large knowledge bases such as Cyc (Liu, Lieberman and Selker 2003). Although handcrafted heuristics remain an important part of the process, it is hoped that the resulting system is more robust, i.e. applicable over a wider range of text genres.

In the following, we describe how we addressed (not solved!) the issues mentioned above while constructing our own system to determine an affect rating for documents. As in most other recent work, affect is treated as running on a linear scale from negative (angry, unhappy) to positive (pleased, happy).

2. **System architecture**

Our affect rating system serves as a component of customer service software, and is used principally in routing email and alerting email recipients to the emotional tone of messages. It can also be used in creating various reports, where it can aid in identifying especially urgent concerns of customers, or especially appreciated products or services.

The system architecture relevant to affect analysis is straightforward. A document is first part-of-speech tagged, and individual rated words (i.e. those on a proprietary list) are identified. Then modifiers such as "very" or "slightly" are detected, and the ratings of words immediately following are recalculated via a modification function. Next, for sentences containing both rated words and negation, syntactic rules are applied to determine whether the negation applies to the rated words, and the sentence ratings are adjusted. Finally, an overall rating is assigned to the document.

3. **Affective word ratings**

The wordlists provided with our application were assembled from those generated by ourselves and colleagues, and the ratings were obtained as averages over a number of raters. The number of rated words is a few thousand. Because the significance of many words depends on context, we asked raters to give their best estimate over a range of contexts, with a bias toward neutrality. Certain words are designated as extreme "cusswords," and their presence above a threshold in a message results in a maximum negative overall rating, regardless of otherwise compensating positively rated words.

In our system, the issues of changing vocabulary and local context are largely handled by allowing our diverse users to change default ratings and add their own rated words. Within the sphere of a particular organization, usage is less ambiguous than in the language as a whole. In some settings, words can even be reliably taken to have the opposite connotation from that in general usage.

To obtain wordlists in foreign languages, we start with translations of rated English words, but also ask native speakers to make the inevitable corrections and additions.

4. Rule construction

Syntactic rules for negation, like word connotations, involve ambiguity. For example, in the sentence "I did not give John a red apple yesterday," the focus of the negation could be any of the words following "not." The intonation that could distinguish the possibilities in speech is not available in a text document. However, this problem is not so severe for our purpose, since oppositely rated words do not frequently appear within the scope of a given negation. Some rules can thus allow the negation to apply to several words within the scope.

The search for rules that might apply to a particular tagged sentence runs from specific to general, with slowly declining confidence scores. The first rule that fits is the one used. If no matching rule is found for a sentence with negation, the score for that sentence is reduced, making it less likely to distort an overall message rating.

The creation of the negation rules is a somewhat tedious chore, especially as it must be repeated for each language, so we have constructed tools to assist with rule development. Scanning corpora consisting of miscellaneous documents assembled from the web, we display in alignment the most common tagged patterns containing negation words. The coverage of the rules varies, but comprises about 96% of such "typically" occurring patterns in the case of English.

5. Evaluation

Based on our experience, the system described performs well on "typical" customer emails. A more rigorous evaluation requires a corpus of rated messages, which we are in the process of establishing. Meanwhile, as an experiment, we tested our method on the collection of movie reviews assembled by Pang, Lee, and Vaithyanathan (available at http://www.cs.cornell.edu/people/pabo/movie-review-data). Using the default wordlists, our system had an accuracy of 63%. This is well below the best result of 83% obtained by Pang et al, who trained various machine learning algorithms on that specific dataset, and slightly below the 66% obtained by Turney (2002), who used a mutual information technique on a different movie review dataset. Our accuracy improved to 73% if we ignored reviews that received a neutral score (within 5% of full scale from the scale midpoint). Significantly, our system performs far better on the positive movie reviews than on the negative ones: the latter are actually more likely to be rated positive than negative! This suggests that there may be strong genre-specific effects, in accord with discussion in the references cited.

6. **Summary**

Based on our experience, we believe it is feasible, with modest effort, to construct a system for affect rating of texts within a particular domain. We have carried this out for multiple languages. Combining methods like ours with unsupervised learning techniques like Turney's, and with real world knowledge along the lines of Liu et al, could increase accuracy over a wider range of genres without the need for labeling or further wordlist creation. Additional improvement will probably require advances in other NLP areas such as word sense disambiguation and discourse understanding.

7. **References**

Kushal, D., Lawrence, S. and Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Proceedings of the Twelfth International Conference on World Wide Web, pp. 519-528. New York:ACM Press.

Liu, H., Lieberman, H. and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. Proceedings of the 2003 International Conference on Intelligent User Interfaces, pp. 125-132. New York:ACM Press.

Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T. (2002). Mining product reputations on the web. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 341-349. New York:ACM Press.

Pang, L., Lee, L. and Vaithyanathan (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 79-86.

Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. 54, 547-77.

Spertus, E. (1997). Smokey: automatic recognition of hostile messages. Proceedings of the Ninth Annual Conference on Innovative Applications of Artificial Intelligence, pp. 1058-1065. Menlo Park:AAAI Press.

Stone, P. J., Dunphy, D. C., Smith, M. S. and Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge:MIT Press.

Subasic, P. and Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. IEEE Transactions on Fuzzy Systems 9, 483-496.

Tong, R. (2001). Detecting and tracking opinions in online discussions. SIGIR 2001 Workshop on Operational Text Classification. http://www.daviddlewis.com/events/otc2001/presentations/otc01-tong-paper.pdf.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp. 417-424.