Domain Specific Affective Classification of Documents

Sara Owsley and **Sanjay Sood** and **Kristian J. Hammond**
Intelligent Information Laboratory
Northwestern University
2133 Sheridan Road
Evanston, IL 60208-3118
sowsley@cs.northwestern.edu

Abstract

In this paper, we describe a set of techniques that can be used to classify weblogs (blogs) by emotional content. Instead of using a general purpose emotional classification strategy, our technique aims to generate domain specific sentiment classifiers that can be used to determine the emotional state of weblogs in that domain.

Keywords

Emotional Analysis, Affect Classification, Blog Processing, Sentiment Analysis.

Introduction

The emergence of Internet weblogs (Blogs) as a primary source for first-person, human experience and opinion has been a driving force in the democratization of media. This force has provided outlets for a varied set of people from all across the globe to share their everyday interactions, personal experiences, and opinions on a wide array of topics.

The creation of search engines specifically tailored for retrieval of blog entries (Google Blog Search 2005) (Technorati 2005) has made this information massively searchable and topically accessible. With access to these large repositories of data, opportunities arise to computationally analyze the content for specific purposes.

For example, companies worldwide spend billions of dollars every year on marketing. A large part of consumer direct marketing involves focus groups evaluating and providing opinions on specific products or services. The existence of publicly available blogs provides another source for these opinions.

In order for these large sets of blog entries to be useful, however, automated evaluation and analysis techniques are critical. Specifically, given the ability to emotionally analyze topically focused text, a software system can provide both quantitative and qualitative evaluation of blogs on that topic.

We have begun exploring the realm of automatic opinion aggregation and classification to make sense of the vast amount of information available through the millions of blogs on the Internet. As part of this goal, we are exploring ways to determine a blogger's opinion on a topic/product, as well as track a blogger's opinion on that topic/product over time.

We envision a system that performs such opinion tracking using a domain specific approach. This system contains two major components. First, the system must determine the domain (e.g., movies, politics, sports, cooking) of an unclassified weblog. Secondly, the system will perform domain based sentiment classification. In this paper, we begin to explore the latter.

We believe it's necessary to use domain specific language to classify the emotional content. Domain specificity is critical in making this system work, since the language used to describe automobiles (sleek, maneuverable, etc.) is different from the language used to describe vacation destinations (relaxing, adventurous, etc.). Previous work in movie review sentiment classification (Pang, Lee, & Vaithyanathan 2002) (Turney 2002) (Turney & Littman 2003) and the large amounts of readily available data makes this domain an appealing starting point.

Previous Work

Others have made attempts at automated affective classification of documents (Ortony, Clore, & Foss 1987) (Glance *et al.* 2005) (Boucouvalas 2002). Some have used WordNet (Miller, Fellbaum, & Miller 1993) to mine sets of affective adjectives (Kamps & Marx 2002) (Hu & Liu 2004). Kamps and Marx scored adjectives on multiple dimensions using a combination of the distances (synonym depth on WordNet) from two defining polar synonyms (like good/bad, passive/active). This method, called the "Semantic Differential", was previously explored by Osgood in his work on "Semantic Space" (Osgood, Succi, & Tennenbaum 1957). A bias towards positive words as well as problems with accuracy in shorter documents (false positives and false negatives) prevail as obstacles for this method.

Previous work has been done in sentiment analysis of movie reviews (Pang, Lee, & Vaithyanathan 2002) (Turney 2002) (Turney & Littman 2003). Pang, et al, used various machine learning techniques, trained by star-rated reviews, to classify reviews as positive or negative. Turney, et al, built an analogous system to classify movie reviews that used co-occurrence between phrases and the words "excellent" and "poor" on the Web. Both systems had promising results.

Method

As with previous methods of movie review sentiment classification, we viewed sentiment analysis as a text categorization problem with two categories, positive and negative (Pang, Lee, & Vaithyanathan 2002). For classification, we decided to use a Naive Bayesian Classifier (NB) because of its robustness and relative ease of implementation.

To train the NB Classifier in the domain of movies, we built a web crawler that collected 10,000 random movie reviews from the Internet Movie Database (IMDB) (The Internet Movie Database 2005). The reviews contained meta-information including a star ranking from 1 to 10 stars (1 is most negative, 10 is most positive). Given the 10,000 movie reviews in our set, we selected a subset of reviews in which the author assigned a star rank of 1 or 10 stars. The final training set contained 1200 1-star movie reviews and 1200 10-star movie reviews for a total of 2400 training documents. While previous systems have been trained on the entire spectrum of reviews, we decided to train on only the extremes (1-star and 10-star reviews) to determine whether classification accuracy could be improved by reducing the training noise and training set size.

Given the collected training corpus of reviews, we chose adjectives as our training feature because adjectives, by definition, are descriptive words and could aide in sentiment detection. We used Brill's Part of Speech Tagger (Brill 1995) to extract the adjectives from each document in the training corpus, and used Porter's stemmer (Porter 1980) to find common roots. We treated each adjective as a feature of the document. Each document was decomposed into a vector representing the number of times each feature occurred. This vector along with known sentiment classification was used to build final training probabilities for the NB Classifier.

A total of 3180 unique adjectives (22638 total occurrences) were observed in the positive data set and 2923 unique adjectives (21885 total occurrences) in the negative set. See table 2 for a small excerpt of positive and negative adjective probabilities based on this training.

Given this training data, the sentiment of a movie review was determined by calculating the maximum NB probability of the movie review *d* being a member of each candidate class *c* (positive or negative), using the following equation:

$$P_{NB}(c|d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

P(c) was derived from a separate, random crawl of 10,000 movie reviews from IMDB to get the relative frequency of positive and negative reviews. We labeled movie reviews with 1 to 4 stars as negative and 7 to 10 stars as positive. *P(neg)* was 0.213 and *P(pos)* was 0.658.

$P(f_i|c)$ is the probability that a feature (an adjective) will appear in a document of the candidate class *c*, using add-one smoothing. To prevent underflow in the product of the probabilities, we calculated a summation of the logarithm of the probabilities. The resulting equation is:

$$P_{NB}(c|d) = P(c) + (\sum_{i=1}^{m} \log(P(f_i|c)^{n_i(d)}))$$

Evaluation

We sought to both evaluate the emotional classifier's performance when classifying documents in the movie review domain, as well compare the contents of our domain specific corpus to a general purpose affective corpus, the ANEW corpus.

The Movie Review Sentiment Classifier

To test our classifier, we created a testing corpus of 5000 movie reviews from a separate, random crawl of IMDB. Each review had a rating from 1 to 10 stars. The entire collected corpus had a star distribution similar to the P(c) probability described above.

The classification accuracy on this testing corpus was 78.06%, which is significantly better than a random choice accuracy of 50%. While these results are not measurably better than previous work in movie review sentiment classification, the results are comparable despite our training set of reviews with 1 and 10 stars only.

The Corpus of Emotional Adjectives

For comparison, we tested our domain specific word corpus (see table 2), against a corpus created by a psychological study. The Affective Norms of English Words (Bradley & Lang 1999) (see table 1) corpus contains 1,034 unique words with affective valence (a scale from unpleasant to pleasant/negative to positive), arousal (a scale from calm to excited), and dominance (a scale from submissive to dominated) scores on a scale of 1 to 9.

We wanted to see how well our system could classify words in the ANEW corpus. Since the valence in the ANEW list most closely relates to positive/negative sentiments, these values were compared for each word in the ANEW corpus. A close correlation between the two would indicate that a general purpose affective lexicon would suffice for classification in this domain, and maybe others. A weak correlation could be interpreted as a need for specialized domain affective classifiers.

Out of the 1,034 unique words in the ANEW list, 386 (37.3% of the ANEW corpus) had a sentiment probability (positive, negative or both) in our training data and could be classified by our tool. Words in the ANEW list with a valence between 1 and 4 were interpreted as negative and those between 6 and 9 as positive. The negative words in the ANEW list were classified by our system as negative with an accuracy of 57.7% and positive words were classified as positive with an accuracy of 64.4%. This evidence suggests that the general purpose affective lexicon classifiers may perform poorly in the domain of movie reviews and perhaps other domains as well.

Conclusion

The use of domain specific corpora for emotional classification of text has very promising results. By understanding and leveraging the fact that people use varied language to describe objects in different domains, we can tune emotional classification engines to increase accuracy and begin

$Word$	$Valence$	$Arousal$	$Dominance$
gloom	1.88	3.83	3.55
glory	7.55	6.02	6.85
hardship	2.45	4.76	4.22
joyful	8.22	5.98	6.6
menace	2.88	5.52	4.98
terrific	8.16	6.23	6.6

Table 1: A sample of words and their scores from the ANEW list, an affective word corpus created from a psychological study.

$Word$	$P(word\|pos)$	$P(word\|neg)$
great	0.02275	0.00635
human	0.00296	0.00091
classic	0.00234	0.00091
bad	0.00472	0.03121
predictable	0.00017	0.00242

Table 2: A sample of words and the probability that they are negative or positive from the movie domain specific affective adjective training set.

building special purpose classifiers for human interest domains. With the current drive to build reliable and scalable blog categorization tools, opportunities will arise to extract the data necessary to build category specific classifiers.

Given the results of our sentiment classifier, we believe that there are additional techniques that we can apply to dramatically increase the accuracy of classification. Term weighting, automatic generation of domain specific stop lists, and feature selection are among the techniques that we are currently investigating.

We believe that this entire system can be applied to other domains, such as consumer goods (using reviews from Amazon), electronics (using reviews from CNET), politicians, music, and others (with reviews available on sites such as http://www.rateitall.com).

References

Boucouvalas, A. C. 2002. *Emerging Communication: Studies on New Technologies and Practices in Communication*. IOS Press. chapter Real Time Text-to-Emotion Engine for Expressive Internet Communications, 305–318.

Bradley, M. M., and Lang, P. J. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical Report C-1, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics* 21:543–565.

Glance, N.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; and Tomokiyo, T. 2005. Analyzing online discussion for marketing intelligence. In *Proceedings of the 14th International Conference on the World Wide Web*, 1172–1173.

Google Blog Search. 2005. http://blogsearch.google.com/.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth international conference on Knowledge Discovery and Data Mining*, 168–177.

Kamps, J., and Marx, M. 2002. Words with attitude. In *Proceedings of the First International Conference on Global WordNet*.

Miller, G. A.; Fellbaum, C.; and Miller, K. J. 1993. Five papers on WordNet. http://www.cogsci.princeton.edu/∼wn.

Ortony, A.; Clore, G. L.; and Foss, M. A. 1987. The referential structure of the affective lexicon. *Cognitive Science* 11:341–362.

Osgood, C. E.; Succi, G. J.; and Tennenbaum, P. H. 1957. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, 79–86.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Technorati. 2005. http://www.technorati.com.

The Internet Movie Database. 2005. http://www.imdb.com.

Turney, P. D., and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21(4):315–346.

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