Nate Agrin
December 13th, 2006

Developing a Flexible Sentiment Analysis Technique for Multiple Domains

Introduction:

Sentiment analysis of blog text, review sites and online forums has been a popular subject for several years in the field of natural language processing. Researchers have shown that several techniques can successfully estimate the opinion polarity of a given text. This project focuses on taking the initial steps towards creating a metric of sentiment for use in enhancing search and retrieval of opinionated content. Three approaches are identified; bag-of-word classification, lexical rule analysis using term expansion, and statistical classification through rule generated features. By comparing and contrasting these methodologies, it is hoped that a robust technique can be developed to quickly estimate a post's general polarity.

Data sets:

Two sentiment oriented data sets were found and used to develop classifiers. A third data set, known as the General Inquirer (GI) was used to in various steps of the classification procedures.

The corpus developed by Bo Pang and Lillian Lee was used as a primary data set [5]. This data is comprised of 2000 movie reviews split evenly into 1000 positive documents and 1000 negative documents.

The second data set, developed by Minqing Hu and Bing Lui, was originally created in an attempt to mine product data and opinions from online review message boards [2]. This data identifies product features and ranks each opinion within a sentence for five different products on a scale from -3 to +3. Sentences may have more than one opinion measurement depending on the editor's discretion.

Data in the General Inquirer set is maintained by Harvard University and is comprised of several hand annotated dictionaries which attempt to place specific words into various descriptive categories [1]. A wide range of affective categories exist, including "positive", "negative", "pain", "pleasure", "yes", and "no". Each word may belong to more than one category, and more than one instance of a word may exist depending on its contextual use.

Methods:

All programming was done using the Python scripting language. Unless otherwise noted, all statistical analysis was completed with the open source Weka Natural Language Processing Toolset [3].

Data

In the Hu data set, the numeric weights were removed, while the overall positive and negative sentiment was retained. Each sentence in a post was tallied for positive and negative sentiment. The sum of all the sentences opinion scores was calculated and used to define an overall sentiment of each post. Using the total score, the reviews were separated into positive and negative domains, where any post with a score larger than zero was positive and any post with a score less than zero was considered negative. For this survey neutral posts with a total score of zero were removed from the pool of posts.

Data from the General Inquirer was used to create a lists of affective words based on the predefined categories “positive”, “negative”, “pain”, “pleasure”, “yes”, “no” and “negation”. For simplicity, each list of affective words was compiled by establishing a priority list of the various categories which followed the following pattern “positive” > “pleasure” > “negative” > “pain” > “negation” > “yes” > “no”. Words were categorized based on this priority list and placed into one of the five categories. Words were not allowed to exist in more than one category.

Statistical Bag Of Words Approach

A baseline for statistical classification was created by first measuring the standard term frequencies within documents. The most common 500 terms found within the first 500 positive and negative documents (for a total of 1,000 documents) of the Pang corpus were then chosen as attributes for a Support Vector Machine (SMO in Weka) classifier and Complement Naive Bayes (CNB) classifier. From the remaining 500 positive and negative documents in the Pang corpus, a testing set was built and used to determine the accuracy of each classifier.

In order to test the cross corpus accuracy of the classifier, another testing set was built using the Hu data set. Again the top terms found in the Pang training set were used to construct the Hu testing set, and then run using the SMO and CNB classifiers.

In an attempt to bias the classifier towards only affective words, the previous tests were rerun using the compiled General Inquirer data set, as a filter. Each term that was found in the General Inquirer data set was recorded and used to construct a term frequency for each document. This feature set was then used to train a SMO and CNB classifier.

Lexical Rules Approach

Using lexical rules, a baseline was created by tokenizing each sentence in every document within the Pang corpus and then testing each token, or word, for its presence within the compiled General Inquirer data set. If the word existed and was associated with a positive sentiment, a +1 rating was applied to the post’s overall polarity score. Similarly, if a word was found to be associated with a negative sentiment within the General Inquirer data set a -1 rating was applied to the post’s overall score. Each post starts with a neutral score of zero, and was considered positive if the final polarity score

was greater than zero, or negative if the overall score was less than zero, in a process similar to that used by Mishne [4].

In an attempt to improve the accuracy of the simple General Inquirer baseline approach, term expansion using Wordnet was employed as in Hu [2]. Each sentence was tagged with a part of speech tagger trained on the Brown review corpus. Any identified adjectives were then checked for their General Inquirer polarity. If they were found, term expansion was not employed. If there were not found, then the term was passed through Wordnet, creating a similarity tree of other words related to the query word. This tree was then compared to the similarity trees for each word in the General Inquirer data until an intersection of the query's tree and a General Inquirer word's tree was found. An assumption was then made that a similarity intersection implied both terms had the same polarity and thus the unknown word was assigned the known word's polarity.

A final attempt to improve the lexical rules based classification accuracy included a sliding window for negations, along with the Wordnet term expansion approach. Negations were identified by their part of speech tags, and if found within five terms of an affective adjective, it was assumed the adjective's polarity was effectively reversed. Thus any positive adjective would be ranked negatively and any negative adjective ranked positively. The score required by a post to be considered positive was increased to greater than five, in an attempt isolate negative posts that might have a very low, yet positive polarity score.

Statistical Classification via Lexical Feature Space Approach

The final approach attempted to use the powerful aspects of the statistical classifier, while separating out specific terms into feature spaces, as a means of creating a corpus independent sentiment classifier. Sentences were tokenized into single words, and the General Inquirer data was used to create a set of features, using its own categories "positive", "negative", "negation", "yes" and "no". Each time a word was found in one of these categories, the total count for that category within a single document was incremented by one.

Each sentence was tagged and each tag was parsed for negations. Every negation found in the tagged set incremented the tag negation count by one. Finally, the tags were chunked to look for implied "inversions" of sentiment, through parsing of higher order sentence structure such as "I should have had a quite a bit of fun." Each instance of an inversion incremented an inversion count for the specific document, by one as well.

1,500 documents from the Pang corpus were used as training data, while 500 Pang documents were compiled as testing data. 200 documents from the Hu corpus were also used to test the accuracy of this new methodology across different corpora.

Discussion:

Figure 1 shows that statistical classification techniques proved to be the most accurate at predicting the sentiment of a document, scoring up to 81% accuracy, when used

within a single corpus. Classification between the Pang and Hu corpora produced significant positive results by classifying up to 68% of the documents correctly, however it's accuracy as compared to classification within its own corpus was notably worse. The differences between inter- and intra-corpus classification methods is most likely the result of the significant differences between the corpora used in this study. The Pang corpus was noticed to have longer, more formally written movie reviews, while the Hu corpus was written using an informal common vernacular. Hu's corpus was also not intended to be used as a generalized "positive" and "negative" domain. Because the methods used to compile the Hu corpus into "positive" and "negative" categories may have incorrectly classified documents, these errors could make using the corpus as a test set inappropriate. It was noted that that CNB classifier performed slightly better in these studies.

Figure 1

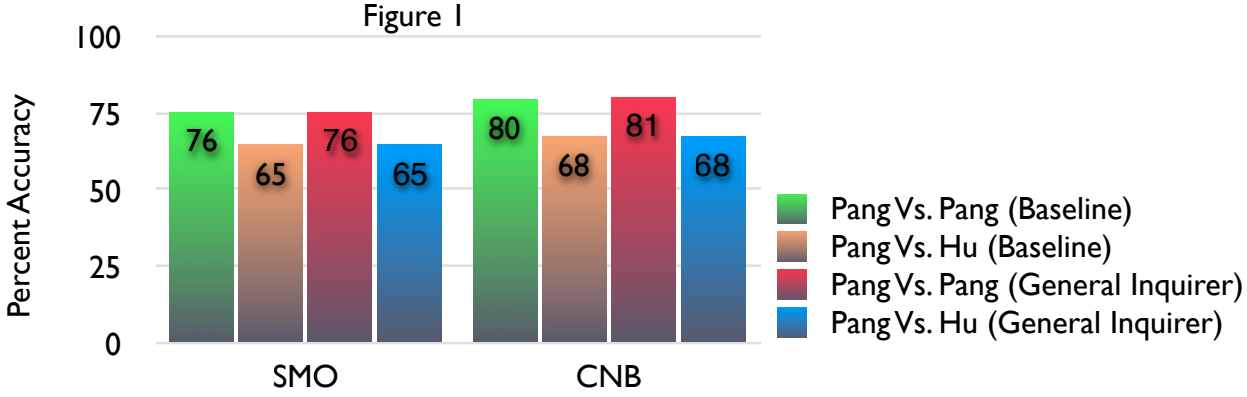

Figure 2

Baseline

	pos	neg
pos	893	726
neg	107	274

Accuracy: 58%

Wordnet

	pos	neg
pos	862	695
neg	102	245

Accuracy: 55%

Wordnet &
Negation w/ positive > 5

	pos	neg
pos	623	290
neg	377	710

Accuracy: 67%

Lexical methods, as described in figure 2, were initially not very accurate at determining the polarity of a document; a standardized baseline approach was only slightly more accurate than random classification. A slight drop from the baseline in accuracy was noted when implementing a Wordnet term expansion process. In order to implement the Wordnet classifier, only adjectives identified through the tagging process were used

to check against the GI data, potentially providing a smaller set of words for which to judge the document's overall polarity, if words were inappropriately tagged.

From the first two lexical experiments preformed, it was noticed that documents tended to be classified as positive. Further inspection of the final polarity scores showed that many negative documents were considered positive while only having a score just slightly above zero. Changing the cutoff value of positive documents to requiring a score of five in the third experiment greatly improved the accuracy of the classifier. Surprisingly, this small change, while only tested on the Pang corpus, was able to classify documents as accurately as in the final approach which used statistical methods on rule based derived features.

Adjusting the positive cutoff in order to increase the lexical classifier's accuracy suggests that this methodology may not be corpus independent, and still relies on close examination of the effects of a classification process on a particular corpus.

Figure 3

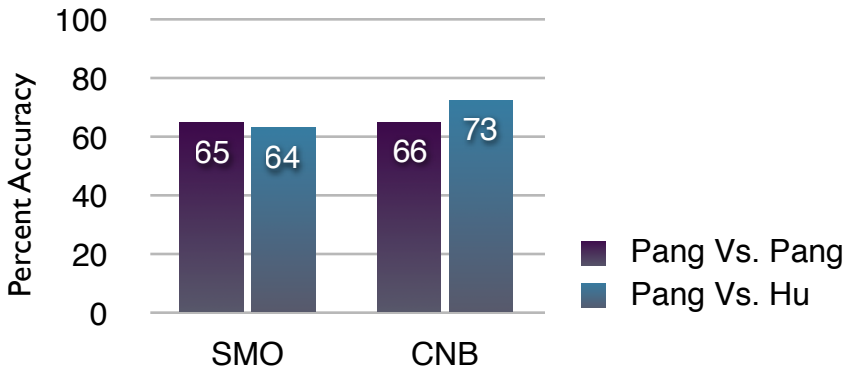

Figure 3 shows that reducing the feature space to features derived from rule based methods, in an attempt to train a classifier to work independently of specific words within a corpus was not significantly more effective than either a rule based or bag-of-words approach. Of interest is the notable increase in accuracy between corpora when using the CNB classifier. It is unclear why across different corpora this approach resulted in higher inter-corpus accuracy than intra-corpus accuracy, suggesting that this methodology should be applied to other corpora to determine if the result is spurious, or meaningful.

Conclusions:

Creating a domain independent sentiment classifier is not a simple task. This evaluation proposed three different approaches and found that each was only capable of accurately classifying documents across domains with a maximum accuracy of approximately 68%. Alternatively, creating a sentiment classifier for a particular domain was capable of classifying documents at an accuracy up to 81%. These results were ob-

tained by using statistical methods and given a sufficient amount of training data. It is suggested therefore that sentiment classification may remain a domain dependent task, dictated by the types of writing used and specific nuances within a document set, best handled through statistical classification methodologies.

Future work to further improve statistical classification would include continued exploration of the rule based feature space for statistical classification. By combining the bag-of-words approach with the features developed through the rules based approach, the accuracy of the classifier might increase by taking into account specific syntactic structure found within sentiment related documents, as well as the frequency of affective terms.

References:

1. General Inquirer. 1997. Data available at: http://www.wjh.harvard.edu/~inquirer/

2. Hu, M., Liu, B. 2004. Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Data available at: http://www.cs.uic.edu/~liub/FBS/FBS.html

3. Ian H. Witten and Eibe Frank. 2005. "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005. Data available at: http://www.cs.waikato.ac.nz/ml/weka/index.html

4. Mishne, G., 2006. “Multiple Ranking Strategies for Opinion Retrieval in Blogs”, The University of Amsterdam at the 2006 TREC Blog Track.

5. Pang, B., Lee, L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the ACL. Data available at: http://www.cs.cornell.edu/people/pabo/movie-review-data