Distinguishing Opinion from News
Katherine Busch

Abstract

Newspapers have separate sections for opinion articles and news articles. The goal of this project is to classify articles as opinion versus news and also to do analysis of the results to figure out the factors that distinguish the two. Preliminary results show that classification is possible with unigram features in an SVM with F1 of .94.

Introduction

This project focuses on subjectivity classification for news articles. Much prior work on subjectivity has focused on distinguishing positive and negative sentiment (for instance, in product reviews) or classifying phrases or clauses as subjective (Liu, 2008). Here we attempt to distinguish entire articles as reporting news or expressing opinion. The task is related but has some key differences. For instance, review-type sentiment analysis often relies on pre-made lexicons or focuses on classifying words as positive or negative (Toprak and Gurevych, 2009; Turney and Littman, 2010; Potts). Words associated with positivity and negativity are not necessarily those associated with editorials and opinion pieces in which authors pose sophisticated arguments about current events, policies, etc. One goal of the project was to gain a lexical understanding of words that can distinguish the two categories, and thus be able to generate a lexicon similar those already existing for sentiment analysis of reviews that would work for articles.

Prior Work

There has been thorough research into document classification."Machine Learning in Automated Text Categorization" (Sebastiani, 2003) provides an overview of work up to 2002. Within the area of subjectivity/sentiment analysis there is also a wide variety of work. Pang and Lee give an overview of the field of subjectivity (Pang and Lee, 2008). Liu defines many different problems within the field including *sentiment and subjectivity classification*:

> (1) classifying an opinionated document as expressing a positive or negative opinion, and (2) classifying a sentence or a clause of the sentence as subjective or objective
>
> (Liu, 2010)

Liu also gives an overview of the field thus far from a teaching perspective. Turney and Littman provide a method for sentiment for particular words based on their context (Turney and Littman, 2003). Yu and Hatzivassiloglou specifically address distinguishing opinion from news using a Naive Bayes classifier and are able to achieve very high results (Yu and Hatzivassiloglou, 2003). However, their method did not generalize to my dataset.

Data

I use two datasets, both consisting of articles from the *New York Times*. The primary dataset consists of 140 articles over the course of the 7 years up to and including 2012. For comparison, I also test on a dataset of articles from October and November 2012 in which news events are covered repeatedly. The data was collected by scraping the *New York Times* website. The first set includes 15 news articles and 5 opinion articles/year arbitrarily selected. The second includes the entire world and United States news sections and entire opinion sections for several days in the past months. The discussion below concerns the long-term set unless otherwise specified.

Results

Dataset: New York Times articles 2006-2012

Learning Algorithm	**F1**
Multinomial NB with Laplace smoothing	.84
SVM: unigram counts	.85
SVM: unigram counts with stemming	.89
SVM: TFIDF	.70
SVM with PoS tags counts (32 features)	.67
SVM with PoS and stemmed unigram	.87
SVM with top 1500 features	**.94**

Table 1: F1 for large time period dataset

Analysis

I focus on the results for the mixed years dataset and only use the small time period dataset for comparison in the *Language results* section. Overall, our classifier achieved high precision and recall for the test set with the best F1 score of .94 using a linear SVM with unigram counts as features, well above the Multinomial Naive Bayes baseline of .84. Below we detail the techniques we tried and where we succeeded.

Feature and classifier selection

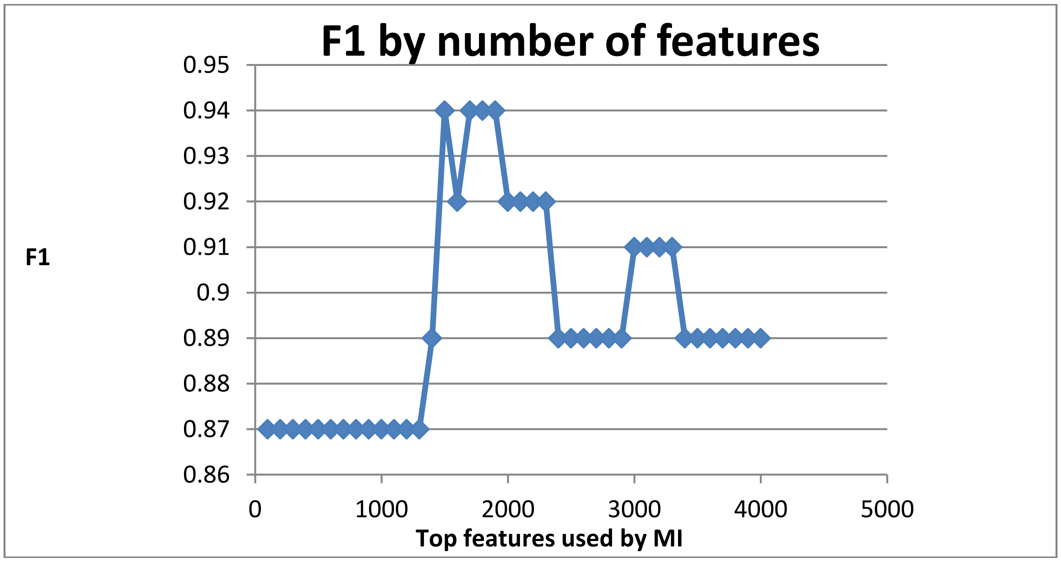

Figure 1: Searching top features (by mutual information score) for optimal number of features

For classifier selection, we tried binomial and multinomial Naive Bayes and SVM with a linear kernel (other kernels discussed below). With initial features, SVM outperformed Naive Bayes with an F1 of .85 compared to .84 for Naive Bayes. While the difference is slight, it widened as we sought to improve the performance as discussed below. Other results suggest than in general SVM outperforms Naive Bayes in text classification, so this difference is expected and we will focus on SVM for the rest of the paper due to its superior performance (Rennie, 2003).

For feature selection, we began with unigram counts using stop-words. With just unigram counts alone, an SVM achieved .85 F1. Adding bigrams and trigams did not improve the classifier at all, and this is likely due to sparsity of data. Sparsity is a common problem with unigram models in which the number of features is much less than the number of training examples (Ng). With unigrams alone, our feature space had 4066 features, yet only 78 training examples. Adding n-grams only increases that space.

Instead we tried several successful techniques for reducing the feature space. With porter stemming, which reduces the feature space by merging words with the same roots, the score increased to .89. Using a mutual information measure of binarized features vectors, we searched the space of number of features in increments of 100, peaking at 1500 features and an F1 of .94. This indicates that the top 1500 words are better for distinguishing opinion from news than the space of all of the features. However, it is interesting to note that just the top 100 features were able to achieve an F1 of .87 which is still very high. After that, the gain in score per feature diminishes greatly, so a classifier interested in efficiency and willing to compromise slightly on correctly could do extremely well in this 100-dimensional feature space.

We also tried using TF-IDF instead of counts. Previous research has suggested that TF-IDF improves scores for text classification (Rennie et al, 2003; Toprak and Gurevych , 2009). We were unable to replicate these results and instead saw F1 decreased to .70. While we do not have a good explanation for why this should be different, usage of stop words and stemming might have helped eliminate words like "the" that would be overcounted. The goal of TF-IDF is to give higher weight to words that occur a lot in a document but little over the corpus. Another theory is that if a news and opinion piece are about the same event, they will have high TF-IDF for words related to that event but that word will not help to distinguish the class. However, the phenomenon is also likely to be a peculiarity of the dataset.

Finally, some work showed that part of speech counts might be effective at subjectivity classification (Toprak and Gurevych, 2009). To test this, we used the counts of part of speech tags from the Penn Treebank tagger. The results were unsuccessful with F1 falling to .67, with articles mostly getting classified as News. Nor did these improve score when used in conjunction with unigram features. The theory behind this is that more adjectives would be used in opinion pieces. That this phenomenon was not observed is probably another difference between newspaper pieces and the traditional reviews that subjectivity research focuses on.

To further improve our results, we tried using both polynomial and Gaussian kernels. While we achieved similar results with these kernels, we were not able to exceed the results from a linear kernel. We believe this is because the initial data was already linearly separable with the exception of a few outliers that the algorithm will not be able to detect.

Indeed, upon looking at the misclassified examples, half were actually reviews of movies, travel destinations, etc, that are not technically classified by the *New York Times* as opinion because they do not appear with the other opinion and editorial pieces. One could argue that the data is mislabeled.

Language-related results

Top-rated by mutual information for short-term dataset:

1. quot
2. year
3. party
4. years
5. israel
6. federal
7. bbc
8. united
9. ms
10. time
11. tax
12. officials
13. city
14. medicaid
15. court
16. women
17. campaign
18. cuts
19. country
20. american

Top-rated by mutual information for long-term dataset:

1. dr
2. report
3. work
4. product
5. percent
6. iraq
7. includ
8. world
9. project
10. secur
11. studi
12. kill
13. told
14. republican
15. street
16. research
17. plan
18. polic
19. program
20. rais

The world that mutual information measurement found to be most informative of category corroborated the hypothesis that traditional sentiment lexicons such as TUD subjective verb lexicon used in Toprak and Gurevych to some would not be as effective for news articles (Toprak and Gurevych, 2009; TUD).

The short term data set as expected includes more words related to specific news events of the last few months--especially politics related ones that were prevalent during the United States election season, such as *campaign, country, party*, and the word *israel* due to the Israeli attach on Gaza. In the short term, particular news pieces are more successful than opinion or news related words in general at distinguishing the categories.

The long term data set, by contrast, included only one word that appeared to be related to a particular event: iraq. Since the Iraq war lasted over the entire period that the dataset was collected from, the presence of the word makes sense. The rest of the words, such as report, work, percent, kill, or polic seem to be clearly connected reporting or opining.

The long-term top 100 features are available at http://katherinebusch.com/op_news_features.txt as a ready-made lexicon for newspaper subjectivity analysis.

Conclusions

The task of distinguishing opinion and news appears to be ones that can be solved with relatively simple tools, much to the credit of the *New York Times*. Prior work in document classification appears to have been effective at this specific classification task. In the future, it would be interesting to explore generalizing the task to different dataset to test whether the lexicon of news/opinion words generated by the model succeeds in classifying articles from other newspapers, news sources, blogs, etc. One could also try using features related to sentence structure. These would be unlikely to improve score but might provide interesting linguistic insights.

References

Liu. "Sentiment Analysis and Subjectivity." Handbook of Natural Language Processing, 2nd Edition. 2010.

Ng. CS229 class notes. CS299.stanford.edu.

Pang and Lee, “Opinion mining and sentiment analysis.” Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.

Potts. "Sentiment Analysis Tutorial." http://sentiment.christopherpotts.net/

Rennie et al. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." 2003.

Sebastiani. "Machine Learning in Automated Text Categorization." ACM computing surveys. 2003.

Toprak and Gurevych. "Document Level Subjectivity Classification Experiments in DEFT’09 Challenge." DEFT'09. 2009.

TUD subject verb lexicon. http://www.ukp.tu-darmstadt.de/data/sentiment-analysis/subjective-verbs-lexicons

Turney and Littman. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association." ACM Transactions on Information Systems, Vol. 21, No. 4, October 2003.

Yu and Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pages 129–136, Sapporo, Japan. 2003.

Libraries used

sklearn, nltk