Report on Political Leaning Classification

Ben Mitchell & Zach Pezzementi
Computer Science Department
Swarthmore College
Swarthmore, PA 19081
{mitchell,zap}@cs.swarthmore.edu

Abstract

The tasks of document classification and sentiment classification have been explored in the literature, but to our knowledge the task of political classification has not. We use a modified form of a document classification algorithm (Hu and Liu, 2004) to classify newspapers as liberal, conservative, or neutral based on their text. By using a cosine similarity metric in our feature space, we were able to achieve distances that separated openly liberal from openly conservative papers. According to the same metric, we found Time and Newsweek to be fairly centrist, as their distances from liberal and conservative papers were about the same, while the Chicago Tribune displayed a distinct liberal bias. This feature space shows promise for further sentiment or document classification work.

1 Introduction

Document classification is the task of grouping a set of documents based on their content, usually into a fixed number of predefined categories. Document classification schemes have been developed for use in specific domains, such as classifying news stories (Yang et al., 1999) or grouping web posted job openings (Cohen and Hirsh, 1998), as well as more generic algorithms designed to work across many domains (Schohn and Cohn, 2000). The classes are usually broad topics, picked in advance (for example, classifying sports articles as being about baseball, football, or basketball). A fairly simple Bayesian bag-of-words model has been shown to be successful in document classification tasks (Baker and McCallum, 1998).

Sentiment classification attempts to group documents according to the sentiment of the author with respect to the subject. Most previous studies have defined the sentiment classification task as integrating aspects of document classification and text summarization (Fei et al., 2004; Hu and Liu, 2004; Pang et al., 2002). The goal is typically to classify each document (often a product review) as being a member of one of two classes, either positive or negative, though attempts at more complex classification schemes have been made(Yi et al., 2003).

We expected the problem of political sentiment classification to require somewhat different techniques from those used in document classification or standard sentiment classification. Firstly, in typical sentiment classification tasks, the text used as input is written specifically to communicate the information the algorithm is trying to extract. A product review, for example, is written with the intention of expressing the sentiment of the reviewer with respect to the product being reviewed. The sentiment we are trying to detect, on the other hand, is not necessarily stated explicitly within the text. Similarly, most document classifiers need only identify the main topic of a passage in order to make their classification decision, whereas we specifically want to avoid distinguishing between articles based primarily on their main topic. To help avoid classifying based on con-

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

Nation	Freq	*National Review*	Freq
dlc	61	guevara	33
un	60	u.n.	29
durbin	42	gannon	26
henry	40	official	26
trotsky	39	chavez	23
falluja	35	pollack	22
guernica	32	kim	20
deutscher	31	ortega	19
nevada	28	mithal	15
women's	25	post-war	14

Table 1: **Top Ten Most Frequent Words Which Occur in Only One Corpus**

tent, we limited our data to articles on a single topic; we chose the United States' war in Iraq as a topic since it was frequently in the news and was also a subject of political contention.

Preliminary tests suggested that unigram probabilities are insufficient for our classification task (see Table 1). The results in this table represent the top ten words in each of two corpora, where words are ranked by number of occurrences, and words that appeared in both corpora were eliminated. To a human observer, there does not appear to be a strong signal of political leaning in these data. For this reason, we used a more complex feature space to do our classification.

2 Procedure

The features we deal with for classifying documents are distributions of association rule confidences as described in (Hu and Liu, 2004). For a given document consisting of a set of words W divided into a set of sentences S, an association rule expresses the likelihood that two separate word phrases X and Y will occur in the same sentence, with an implication that the presence of X causes Y to appear. The rule is defined $X \rightarrow Y$ where $X \subset W$, $Y \subset W$ and $X \cap Y = \emptyset$. That is, both X and Y are word phrases that do not overlap. For our purposes, X and Y are always single words. Two statistics, support s and confidence c, are calculated for each possible word association (every pair of words which occur together in at least one sentence). Support is a measurement of the number of times we see a place in the text where the two words could be associated, and it is defined as the percent of sentences in the corpus that contain either X or Y, $\frac{occ(X \cup Y)}{|S|}$ where $occ(w)$ is the number of sentences containing w. Confidence is then a measurement of how strongly we believe the presence of X causes the presence of Y, and it is measured as the percent of sentences containing X which also contain Y, $\frac{occ(X \wedge Y)}{occ(X)}$.

By imposing thresholds on both c and s (*c-thresh* and *s-thresh*), we select for a given document a number of association rules which both occur somewhat frequently (high support) and have fairly strong causality (high confidence). We further filter these rules by requiring that the "term-sentence frequency" of the second term in the rule, Y, be smaller than a third threshold, *t-thresh*. The term-sentence frequency of a word is defined as the number of sentences containing that word divided by the total number of sentences in the document. This restriction eliminates unimportant rules on very common words like "the" and "of," which would otherwise have very high confidence. The particular values used for these thresholds were *s-thresh*=0.01, *c-thresh*=0.1, and *t-thresh*=0.2. The number of association rules which pass this final threshold define the length of our feature vector for a given document. Some sample rules are given in Table 3. To compare two documents, we use one of several methods to calculate the distance between the feature vectors for the documents. Few rules in a given document's vector occur in other documents' rule-sets as well, so the vectors tend to be fairly distant in the feature space. This means that actual similarity scores will be low, but by comparing relative distances between various publication pairs, we can establish which other publications are more similar to a each other, and which are less.

Our corpora are built from articles obtained via Infotrac both from publications with open political leanings and from those who claim to be balanced or impartial, as shown in Table 2. We restrict our search to articles covering the war in Iraq to minimize variation in the data due solely to topic. This is accomplished by searching the full text for articles containing both "Iraq" and "war."

We calculate distances between articles in the feature space to determine similarity. Distances within

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

Corpus Name	Publication	Size
Liberal		
AProspect	American Prospect	171
Nation	The Nation	110
Nation2	The Nation	102
WashMonth	Washington Monthly	215
Conservative		
Review	The National Review	67
Economist	The Economist	91
Economist2	The Economist	80
WashTimes	The Washington Times	61
"Impartial"		
Time	Time	117
Newsweek	Newsweek	106
ChicTrib	The Chicago Tribune	92
ChicTribBig	The Chicago Tribune	288

Table 2: **Corpus Naming Conventions with Sizes, in thousands of sentences**

humble	$\rightarrow$	bush
exhausted	$\rightarrow$	has
humble	$\rightarrow$	foreign
rare	$\rightarrow$	or
multiply	$\rightarrow$	but
avert	$\rightarrow$	be
attached	$\rightarrow$	be
136000	$\rightarrow$	by
hat	$\rightarrow$	up
beacon	$\rightarrow$	it
instruments	$\rightarrow$	other
omar	$\rightarrow$	an
resounding	$\rightarrow$	not
fortune	$\rightarrow$	his
fails	$\rightarrow$	he

Table 3: **Sample High Scoring Rules from the Intersection of Review and Economist**

the feature space are calculated by one of three distance metrics. Simple cosine similarity is the first. Since $A \cdot B = |A||B|\cos(\theta)$, the cosine of the angle between two feature vectors can be found by computing the dot product of the vectors and dividing by the sum of their lengths; this value can be used as a measure of similarity. We also calculate binary cosine similarity, which is found by converting each non-zero value of the vectors to a 1 and then finding cosine similarity. Our third metric is Euclidean distance in the feature space; we tried using both normalized and un-normalized vectors. The results in Table 9 were generated using the un-normalized vectors, but there did not appear to be a noticeable difference in political-leaning correlation between the two methods.

3 Results

We did several experiments, and the results for each of them are presented here. The naming conventions shown in Table 2 are used to represent our corpora in further figures.

Tables 5, 6, 7 show the data from the cosine similarity comparisons arranged for ease of readability, along with some numerical analyses of those data. For each corpus, the first column shows the publication to which it is being compared. The second column shows that publication's political leaning. The third shows the raw cosine similarity value calculated for those two corpora. The mean and standard deviation over all pairings were calculated, and the fourth column contains the number of standard deviations between that mean and the number in the third column. These tables are sorted according to raw cosine similarity, such that publications more similar to a given corpus appear higher up. The raw data from which these tables are derived is included in Table 8. The raw data for Euclidean distance and binary cosine distance was included in Tables 9 and 10 respectively, but no further analysis was conducted on them since there appeared to be no interesting correlations.

Two corpora were constructed from Chicago Tribune articles, ChicTrib and ChicTribBig. The big corpus contains three times as many articles as the other, and includes *all* of the articles that make up the small one. Table 4 shows cosine similarities of both the small and large corpora to all other corpora. The table is sorted on the second column, and the last column of that table shows the difference in similarity between the second and third columns of that row. This table demonstrates the effect of increased corpus size on cosine similarity. The corpora Nation2 and Economist2 were second corpora taken from The Nation and The Economist respec-

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

	ChicTrib	ChicTribBig	Difference
Time	0.0502	0.0508	+0.0006
WashMonth	0.0492	0.0546	+0.0054
Newsweek	0.0482	0.0508	+0.0026
AProspect	0.0481	0.0530	+0.0049
Nation2	0.0478	0.0502	+0.0024
Nation	0.0459	0.0475	+0.0016
Economist	0.0443	0.0440	-0.0003
Review	0.0442	0.0437	-0.0005
WashTimes	0.0428	0.0366	-0.0062
Economist2	0.0402	0.0395	-0.0007

Table 4: **Cosine Similarities for the Small and Large Chicago Tribune Corpora, sorted from top to bottom by ChicTrib similarity score**

tively so that we could test the similarity of different articles from the same publication. As one might expect, these split corpora were more similar to each other than they were to corpora of other publications. These two "second" corpora were composed of the same number of articles as the "first" corpora, and the first and second corpora did not contain any of the same articles and contained the same number of articles. The second corpora were not included in the larger tables, but the second corpora have similarity scores to other publications comparable to those of the equivalent first corpora.

4 Discussion

While the actual numbers returned by the cosine similarity metric are very small, what we are interested in is the relationships between the numbers. The reason that all the similarity scores are so low is that our rules are generated on a per-document basis; this means that each document is likely to generate many rules which are not generated by any other document in the corpus. Because of this, the document vectors tend to have many "dimensions" (each corresponding to a rule) in which there will never be any overlap. It is this sparsity of rules that are common to multiple publications that causes the similarity scores to be so low. We could determine which rules were generated by only a single corpus and thrown them away, but this process would require our method to deal with all our corpora at once, and for this experiment we wanted to use a method that worked explicitly with only two corpora at a time. Using only two corpora means that a new corpus can be analyzed and compared to any number of existing corpora with relatively little work; working with all the corpora at once would force us to re-analyze every corpus each time we wanted add a new one.

It is therefore not a problem that all our similarity scores seem very small. What is important is the differences between those scores, and how those differences correspond to the differences between the publications perceived by humans. As shown in Tables 5, 6, and 7, the cosine similarity metric gives results that roughly correspond to the "desired" values. Binary cosine similarity and Euclidean distance do not appear to give as meaningful results; there is simply no correlation between political leaning and similarity score (see Tables 8, 9, and 10 for the raw data). The cosine similarity metric gives exclusively higher similarities between publications which are openly liberal than it does between openly liberal and openly conservative publications. The same is not true for the openly conservative publications; similarity between conservative publications is not significantly higher than similarity between conservative and liberal publications. The Washington Times in particular has low similarity to all other publications. This dissimilarity may indicate that it is written in a different style, or that it represents a distinct political category, but it most likely indicates a data scarcity problem, since this was our smallest corpus. As shown in Table 2, the conservative corpora for some reason were all smaller than the liberal corpora, at least in terms of number of sentences, despite the fact that all corpora except ChicTribBig contained exactly 80 articles. This is probably at least part of the reason that the conservative publications have lower similarity to each other; with smaller corpora, there are likely to be fewer rules that overlap.

It is also interesting that the Economist showed up as being closer to the liberal publications than the conservative ones; despite our original label of the publication as "conservative," further investigation has revealed that parts of it are generally considered to be liberal. Had our metric not indicated this to begin with, we would not have known to re-examine our label.

The purportedly impartial publications tested

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

were Time, Newsweek, and the Chicago Tribune. The Chicago Tribune exhibits larger similarities to liberal than to conservative publications. Time and Newsweek, however, both appear fairly balanced in their similarities. The apparently liberal slant of the Chicago Tribune may be in part due to the fact that the liberal corpora contained somewhat longer articles on average. However, tripling the size of the Chicago Tribune corpus makes it more similar to the liberal publications and less similar to some conservative publications, indicating that similarity is not merely a function of corpus size (see Table 4). Additionally, the original Chicago Tribune corpus was the smallest of the three impartial corpora (see Table 2), so the fact that it came out as more liberal goes counter to the trend of bigger being equated with more liberal.

5 Conclusions & Future Work

The method outlined in this paper seems to provide at least some ability to rank the similarity of publications, and the similarities it reports correspond with the political agendas that human readers ascribe to those publications. While these results are encouraging, there is still much work to be done in the area of political sentiment classification.

In the future, we would like to analyze publications which claimed to be impartial but are widely thought to have a political leaning, such as the New York Times, the Wall Street Journal, and the Washington Post. Comparison between these publications and publications with known leanings would be interesting.

We would like to test more and larger corpora, and try to find better values for our constants, possibly by training them using machine learning techniques. We would also like to do more statistical analysis on the results of those tests. This analysis would help to demonstrate more clearly the utility of our method. We especially would like to get more data from the Washington Times, since the WashTimes corpus had very low similarity scores to all of the other publications in the corpus. More experimentation is needed to determine why this is the case, but we did not have a large enough corpus to split that corpus in half and do a self-similarity test, which would be the first test we would do.

It is important to note that our algorithm does nothing to specifically isolate features relating to politics. The fact that the resulting feature space seems able to separate liberal publications from conservative ones may therefore come as some surprise. This result is probably due primarily to the fact that all of the articles dealt with the same general subject-matter. If this had not been the case, it is doubtful that similar results would be obtained, simply because the data would be to scarce for political leaning to dominate article topic. Table 3 indicates that the information captured by our features relates primarily to topic and writing style. The rules generated surprisingly do not look much more meaningful to a human than those in Table 1, but our results show it to be nonetheless sufficient for the task of political leaning classification. Adding further processing that does specifically address politics could produce even better results. One such modification could be to learn a set of words which could be considered important to the domain, such as "politically-charged words". Rules containing those words could be weighted more heavily for intersection in order to focus classification to that domain. Similar modifications could be made to focus on domains other than politics instead, making this technique one of general use in any classification task.

References

L. Douglas Baker and Andrew Kachites McCallum. 1998. Distributional clustering of words for text classification. In *Proceedings of 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98)*.

William W. Cohen and Haym Hirsh. 1998. Joins that generalize: text classification using WHIRL. In Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors, *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173, New York, US. AAAI Press, Menlo Park, US.

Zhongchao Fei, Jian Liu, and Gengfeng Wu. 2004. Sentiment classification using phrase patterns. In *The Fourth International Conference on Computer and Information Technology (CIT'04)*, pages 79–86, Wuhuan, China, Sept.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of Nine-*

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

AProspect (L)				Nation (L)				WashMonth (L)			
Name	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*
WashMonth	*L*	0.0623	+2.460	AProspect	*L*	0.0558	+1.390	AProspect	*L*	0.0623	+2.460
Nation	*L*	0.0558	+1.390	WashMonth	*L*	0.0546	+1.190	Time	*I*	0.0547	+1.210
Time	*I*	0.0510	+1.210	Economist	*C*	0.0497	+0.384	Nation	*L*	0.0546	+1.190
Newsweek	*I*	0.0502	+0.466	Review	*C*	0.0494	+0.334	Newsweek	*I*	0.0528	+0.894
Review	*C*	0.0501	+0.450	Newsweek	*I*	0.0479	+0.087	Review	*C*	0.0503	+0.483
Economist	*C*	0.0491	+0.285	ChicTrib	*I*	0.0459	-0.242	Economist	*C*	0.0496	+0.367
ChicTrib	*I*	0.0482	+0.137	Time	*I*	0.0455	-0.308	ChicTrib	*I*	0.0493	+0.318
WashTimes	*C*	0.0370	-1.708	WashTimes	*C*	0.0342	-2.169	WashTimes	*C*	0.0364	-1.807

Table 5: **Similarity of Liberal Publications to All Publications, sorted by similarity**
P is political leaning. Raw is cosine similarity. Scaled is number of standard deviations from the mean.

Review (C)				WashTimes (C)				Economist (C?)			
Name	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*
Time	*I*	0.0518	+0.730	ChicTrib	*I*	0.0429	-0.736	Nation	*L*	0.0497	+0.384
Newsweek	*I*	0.0507	+0.546	Newsweek	*I*	0.0413	-1.000	WashMonth	*L*	0.0496	+0.367
WashMonth	*L*	0.0503	+0.483	Time	*I*	0.0411	-1.032	AProspect	*L*	0.0491	+0.285
AProspect	*L*	0.0501	+0.450	Review	*C*	0.0376	-1.609	Review	*C*	0.0476	+0.038
Nation	*L*	0.0494	+0.334	AProspect	*L*	0.0370	-1.708	Time	*I*	0.0454	-0.324
Economist	*C*	0.0476	+0.038	Economist	*C*	0.0368	-1.741	ChicTrib	*I*	0.0443	-0.506
ChicTrib	*I*	0.0443	-0.505	WashMonth	*L*	0.0364	-1.807	Newsweek	*I*	0.0442	-0.522
WashTimes	*C*	0.0376	-1.609	Nation	*L*	0.0342	-2.169	WashTimes	*C*	0.0368	-1.741

Table 6: **Similarity of Conservative Publications to All Publications, sorted by similarity**
P is political leaning. Raw is cosine similarity. Scaled is number of standard deviations from the mean.

Time (I)				Newsweek (I)				ChicTrib (I)			
Name	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*	*Name*	*P*	*Raw*	*Scaled*
Newsweek	*I*	0.0548	+1.223	Time	*I*	0.0548	+1.224	Time	*I*	0.0503	+0.483
WashMonth	*L*	0.0547	+1.207	WashMonth	*L*	0.0528	+0.894	WashMonth	*L*	0.0492	+0.318
Review	*C*	0.0518	+0.730	Review	*C*	0.0507	+0.548	Newsweek	*I*	0.0482	+0.137
AProspect	*L*	0.0510	+0.598	AProspect	*L*	0.0502	+0.466	AProspect	*L*	0.0481	+0.137
ChicTrib	*I*	0.0503	+0.482	ChicTrib	*I*	0.0482	+0.137	Nation	*L*	0.0459	-0.242
Nation	*L*	0.0455	-0.308	Nation	*L*	0.0479	+0.087	Economist	*C*	0.0443	-0.506
Economist	*C*	0.0454	-0.324	Economist	*C*	0.0442	-0.522	Review	*C*	0.0442	-0.506
WashTimes	*C*	0.0411	-1.032	WashTimes	*C*	0.0413	-1.000	WashTimes	*C*	0.0402	-0.736

Table 7: **Similarity of Impartial Publications to All Publications, sorted by similarity**
P is political leaning. Raw is cosine similarity. Scaled is number of standard deviations from the mean.

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

	AProspect	ChicTrib	Economist	Nation	Newsweek	Time	WashMonth	WashTimes	Review
AProspect	1.0000	0.0482	0.0491	0.0558	0.0502	0.0510	0.0623	0.0370	0.0501
ChicTrib	0.0482	1.0000	0.0443	0.0459	0.0482	0.0503	0.0493	0.0429	0.0443
Economist	0.0491	0.0443	1.0000	0.0497	0.0442	0.0454	0.0496	0.0368	0.0476
Nation	0.0558	0.0459	0.0497	1.0000	0.0479	0.0455	0.0546	0.0342	0.0494
Newsweek	0.0502	0.0482	0.0442	0.0479	1.0000	0.0548	0.0528	0.0413	0.0507
Time	0.0510	0.0503	0.0454	0.0455	0.0548	1.0000	0.0547	0.0411	0.0518
WashMonth	0.0623	0.0493	0.0496	0.0546	0.0528	0.0547	1.0000	0.0364	0.0503
WashTimes	0.0370	0.0429	0.0368	0.0342	0.0413	0.0411	0.0364	1.0000	0.0376
Review	0.0501	0.0443	0.0476	0.0494	0.0507	0.0518	0.0503	0.0376	1.0000

Table 8: **Simple Cosine Similarity**

	AProspect	ChicTrib	Economist	Nation	Newsweek	Time	WashMonth	WashTimes	Review
AProspect	0.0000	216.6922	212.5889	242.6386	234.4846	229.0122	286.4664	182.5949	211.1575
ChicTrib	273.4382	0.0000	216.0229	246.8631	237.6160	232.0051	290.7928	183.7058	214.8406
Economist	273.0869	219.9037	0.0000	246.6635	237.8007	232.3097	290.4779	184.5227	214.2705
Nation	271.4702	218.7671	214.2648	0.0000	236.5015	231.3484	289.3130	183.9661	212.8550
Newsweek	272.1748	218.4070	215.1057	245.6556	0.0000	230.4619	289.3109	183.1170	213.0973
Time	272.2653	218.5251	215.2385	246.1693	235.9112	0.0000	289.3462	183.4231	213.1218
WashMonth	267.2290	215.4656	211.6379	241.6557	233.1303	227.3656	0.0000	181.7377	210.0709
WashTimes	274.7550	221.4264	217.8855	248.9037	239.6326	234.0268	292.0342	0.0000	216.8134
Review	274.1146	220.5114	216.4998	247.5112	238.1333	232.5915	291.3607	184.8660	0.0000

Table 9: **Euclidean Distance in Feature Space**

	AProspect	ChicTrib	Economist	Nation	Newsweek	Time	WashMonth	WashTimes	Review
AProspect	1.0000	0.1301	0.1321	0.1477	0.1372	0.1339	0.1651	0.0952	0.1328
ChicTrib	0.1301	1.0000	0.1195	0.1228	0.1325	0.1385	0.1330	0.1042	0.1205
Economist	0.1321	0.1195	1.0000	0.1343	0.1211	0.1205	0.1325	0.0943	0.1277
Nation	0.1477	0.1228	0.1343	1.0000	0.1297	0.1229	0.1452	0.0883	0.1333
Newsweek	0.1372	0.1325	0.1211	0.1297	1.0000	0.1467	0.1407	0.1041	0.1351
Time	0.1339	0.1385	0.1205	0.1229	0.1467	1.0000	0.1436	0.1045	0.1349
WashMonth	0.1651	0.1330	0.1325	0.1452	0.1407	0.1436	1.0000	0.0960	0.1304
WashTimes	0.0952	0.1042	0.0943	0.0883	0.1041	0.1045	0.0960	1.0000	0.0955
Review	0.1328	0.1205	0.1277	0.1333	0.1351	0.1349	0.1304	0.0955	1.0000

Table 10: **Binary-valued Cosine Similarity**

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College

teenth National Conference on Artificial Intelligence (AAAI-2004), San Jose, USA, July.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, July.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proc 17th International Conf. on Machine Learning*, pages 839–846. Morgan Kaufmann, San Francisco, CA.

Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. 1999. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, July.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, Florida.

Appeared in: *Proceedings of the Class of 2005 Senior Conference*, pages 34–41
Computer Science Department, Swarthmore College