Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection

Kathleen T. Durant and Michael D. Smith

Harvard University, Harvard School of Engineering and Applied Sciences,
Cambridge MA, USA
{Kathleen,Smith}@eecs.harvard.edu

Abstract. As the number of web logs dramatically grows, readers are turning to them as an important source of information. Automatic techniques that identify the political sentiment of web log posts will help bloggers categorize and filter this exploding information source. In this paper we illustrate the effectiveness of supervised learning for sentiment classification on web log posts. We show that a Naïve Bayes classifier coupled with a forward feature selection technique can on average correctly predict a posting's sentiment 89.77% of the time with a standard deviation of 3.01. It significantly outperforms Support Vector Machines at the 95% confidence level with a confidence interval of [1.5, 2.7]. The feature selection technique provides on average an 11.84% and a 12.18% increase for Naïve Bayes and Support Vector Machines results respectively. Previous sentiment classification research achieved an 81% accuracy using Naïve Bayes and 82.9% using SVMs on a movie domain corpus.

Keywords: Sentiment Classification, Blogs, Web Logs, Naïve Bayes, Support Vector Machines, WEKA, feature selection.

1 Introduction

In December 2004, a Gallup Poll reported that over the last two years the only news and related information source showing an increase in daily use was the Internet. Every other news source decreased, and local TV news, local newspapers and network news magazine shows reached new lows. The percentage of Americans getting their news on the Internet every day has increased in Gallup polls from 3% in 1995 to 20% in 2004 [2]. Out of the 94 million Americans using the Internet in September 2005, 46% of them use the Internet daily to read news. It is the third most popular activity on the Internet, surpassed only by ubiquitous activities such as processing email and using a search engine [19].

The number of web logs, also referred to as *blogs*, has increased dramatically in the last few years. An estimated 59.6 million blogs now exist in cyberspace, up from just 100,000 in 2002 [6]. According to Technorati, an authority on blogs, the number of web logs doubles every 6 months with 75,000 new web logs coming into existence every day. The daily posting volume of web log posts is 1.2 million or 18 posts a

O. Nasraoui et al. (Eds.): WebKDD 2006, LNAI 4811, pp. 187–206, 2007.

second. In November 2004, a Pew Poll reported the number of readers accessing information on web logs had increased by 58% over the course of the year [4]. 10% of all Internet users either have a web log or have posted their opinion to a newsgroup or some other online journal. In February 2004, 17% of the Internet users had used the Internet to read someone else's web log; by September 2005, that figure has increased to 27% [16, 17]. In February 2004, 5% of the polled Internet users had used the Internet to create a web log; by September 2005, that figure has jumped to 9% [16, 17]. Using web logs to share ideas and opinions is growing rapidly in popularity and has become an integral part of our online culture.

Web logs provide a mechanism for people to express their ideas and opinions with the world. They allow a writer to share his first-hand experience, thoughts and opinions with anyone in the world that has access to the Internet. The compendium of web logs can be viewed as a plethora of people's opinions. Our research applies sentiment classification to the voluminous collection of opinions found in web logs. Sentiment classification is the ability to judge a passage of text as positive or negative given a particular domain or topic. More specifically, sentiment classification is the ability to label a passage according to its general sentiment $p \in \{-1, 1\}$, where -1 represents unfavorable and 1 represents a favorable description. It divides a collection of opinions into two opposing camps.

We limit our web logs to political web logs; this is a new domain area for sentiment classification research. Previous sentiment classification studies used news articles as its domain [20, 10, 7]. Others used movie reviews [10, 1, 14, 15]. Nasukawa and Yi used camera reviews as their domain [13], and Turney and Littman's corpus was composed of 410 reviews from Epinions randomly sampled from four different domains: automobiles, banks, movies and travel destinations [22]. Das and Chen's research was applied to Yahoo's stock message boards [3].

We believe political web log posts to have different characteristics than the domains in previous studies. Web logs are highly opinionated and rich in sentiment. Predicting the sentiment of a political web post (i.e., predicting that the post came from a liberal or conservative blogger) is more difficult than predicting sentiment of traditional text (e.g., newspaper articles). Nonprofessional writers usually author web logs; the writing takes on a less formal conversational style of documentation. The language used in web logs is quite rich and has many forms of speech such as cynicism and sarcasm. Many times the complete concept of a post can only be determined by the interplay of the text and a picture posted with the text. Other times the sarcasm is so heavy, readers misinterpret the meaning of a post. Hyperlinks also play an important role in the meaning of a web log post. Most web logs contain many hyperlinks; enabling a reader to follow the evolution of a topic from web log to web log. The information from the hyperlinks often enhances the meaning of a post. Our domain can be characterized quite differently than traditional prose and even other online opinionated data; yet we show that a standard machine learning technique perform almost as well in our domain as in other domains and if coupled with a feature selection algorithm can surpass previous results.

We have chosen to create a topic-specific corpus. Our topic is people's opinion on President George W. Bush's management of the Iraq War. Corpuses from previous studies are only domain specific not topic specific [1, 7, 10, 13, 14, 15, 22]. Engström showed machine learning classification to be highly topic-dependent [5]. If given a

topic-specific corpus a machine classifier takes advantage of topic-specific terms and in general produces higher results than if given a nonspecific topic corpus. However, we found an opposite result. Our classifiers trained on our topic specific data using the same standard feature set representation performed slightly worse than a classifier trained on a nonspecific topic corpus [14]. We believe this degradation is due to the characteristics of our web log corpus.

The ability to judge sentiment would be extremely useful when applied to the vast number of opinions found in the growing number of on-line documents such as web logs and editorial pages of news feeds. Predicting and tagging sentiment of a web log post could improve the process of web logging. It could help organize the information and allow users to find and react to opposite or similar opinions thus improving and simplifying the process of sharing and discussing opinions in web logs. In this paper we investigate three aspects of our web log corpus that need to be understood in order to pre-tag the sentiment of web log posts: applicable machine learning techniques, feature selection, and class constituency. We recognize time as an influential aspect of our data and use a simple segmentation scheme but do not investigate other solutions.

We chose to partition our slightly greater than two years of data by the month; thus creating twenty-five partitions. We predict the sentiment of political web posts for each of the 25 different time segments. We believe our data and many of our terms to be time-specific so we keep our data time-ordered. We chose our time interval to be a month because we needed an interval large enough to ensure enough postings to create good-sized datasets yet small enough to limit the number of events discussed within the interval.

We vary dataset creation along two dimensions: class constituency and feature set collection. We also investigate the use of different machine learning techniques such as Naïve Bayes and Support Vector Machines. We wish to determine if existing technology can be successfully applied in our domain. Since we wish to take advantage of all our data, we measure the accuracy of different datasets that consist of balanced and imbalanced categorical compositions. In our first collection we gather as many posts as we can from the web. This approach led to an imbalanced category makeup within our datasets. This imbalance is expected, since the topic may be discussed more ardently in one camp than the other. One camp could be inflamed on a topic; while the other camp ignores the topic. Our second collection balances the constituency of our datasets by randomly discarding posts of the majority class, the class that outnumbered the other class. This approach led to smaller datasets. Smaller datasets tend to produce lower accuracies than larger datasets; however we show balanced datasets produce similar yet unbiased accuracy results. We then considered three different approaches to feature selection. Our first approach limits the features to the terms occurring at least five times within the corpus, a representation used in a previous study [14]. We then added features found within log posts for the current month but were not part of the dataset, yielding on average feature sets 1.75 times larger. The added features did not improve the accuracy of our datasets. Lastly, we applied a forward search feature selection algorithm to determine our features; this technique drastically decreased the number of features. It also improved our results significantly; on average an 11.84% and a 12.18% increase for Naïve Bayes and Support Vector Machines respectively.

2 Previous Work in Sentiment Classification

Previous work can be categorized by the approach used to perform sentiment classification. The knowledge-based approach uses linguistic models or some other form of knowledge to glean insight into the sentiment of a passage. Later approaches apply statistical or machine learning techniques for achieving sentiment classification. A brief history of both approaches follows.

2.1 Knowledge-Based Sentiment Classification

Both Hearst [8] and Sack [20] categorized the sentiment of entire documents based on cognitive linguistics models. Other researchers such as Huettner and Subasic [10], Das and Chen [3], and Tong [20] manually or semi-manually constructed a discriminate word lexicon to help categorize the sentiment of a passage. Hatzivassiloglou and McKeown [7], and Turney and Littman [22] chose to classify the orientation of words rather than a total passage. They used the semantic orientation of individual words or phrases to determine the semantic orientation of the containing passage. They preselected a set of seed words or applied linguistic heuristics in order to classify the sentiment of a passage. Beineke, Hastie and Vaithyanathan extend Turney and Littman's research using a pseudo-supervised approach [1]. They address the problem of the limited number of labeled data by using both labeled and unlabeled data. They defined anchors of sentiment as pairs of words that co-occur frequently and support a positive or negative sentiment. Other words found to occur more frequently with the anchor words are then chosen to be anchor words. They use the anchor words as their feature set and apply a Naïve Bayes classifier to the dataset.

Nasukawa and Yi [13] take a completely different approach to sentiment analysis. They see a topic as an item containing many different parts or features. They wish to identify the sentences that contain opinions concerning the features of the topic. Sentiment analysis involves the identification of sentiment expressions, polarity and strength of the expression, and their relationship to the subject. They choose a particular topic of interest and manually define a sentiment lexicon for identification. The classification of each review was manually determined by a judge rather than the author of the review. They believe this approach provides not just a sentiment class but an analysis of the opinions found within a review. This approach is useful when measuring customer satisfaction of a particular product. It allows a product to be reviewed as a sum of its parts. Many consumers update on-line product web logs; being able to organize and sort positive and negative comments benefits the supplying corporation of a product as well as consumers.

2.2 Statistical Sentiment Classification

Pang, Lee, and Vaithyanathan have successfully applied standard machine learning techniques to a database of movie reviews [14]. They chose to apply Naïve Bayes, Maximum Entropy and Support Vector Machines to a domain specific corpus of movie reviews. They represented the reviews in eight different formats, the simplest being a unigram representation. The accuracy of their most successful representation, the unigram feature set representation, and their most successful machine learning

induction method, Support Vector Machines, produced an accuracy of 82.9%. Their Naïve Bayes classifier with a unigram feature set representation achieved an accuracy of 81.0%. They continued their research by defining a scheme that addresses the nature of a review. They argue a review consists of both objective and subjective sentences, where the objective sentences describe the plot of the movie and the subjective sentences expresses the reviewer's opinion of the story. They created extracts from the reviews that contained the sentences identified as the most opinionated. They achieved some success in this approach creating extracts 60% the size of the original review with accuracy better than or at least as accurate as the accuracy of the full text review [15].

3 From Blogs to Datasets

The website, themoderatevoice.com, is a political web log that lists and categorizes over 250 web logs as *left voices*, *right voices* or *moderate voices*. The list was created by the journalist Joe Gandelman, who classifies himself as a political moderate. Gandelman's categorization of each blog is the information we attempt to predict. We allow postings from a blog to inherit the categorization of the blog and attempt to classify a post as originating from a left voice or a right voice.

We harvested the posts from the left-voice and right-voice blogs for the time period of March 2003 to March 2005. We apply a topic selection filter over the posts. Our filter identifies the posts that contain our specific topic from the selected posts. The following sections discuss the details of the posts collected to create our dataset of political blogs. This discussion is then followed by a description of our chosen feature set representation and values.

3.1 A Description of the Web Data

Out of the 99 left-voice blogs and the 85 right-voice blogs listed in March 2005 on themoderatevoice.com, 84 left-voice blogs and 76 right-voice blogs were included within our study. The other 24 blogs were eliminated because they were political cartoons, lacked archives, were broken links, or were an online magazine that contained no original posts. For a complete list of the contributing web logs please refer to Appendix A. The total size of the right-voices' web files is slightly less than 775 Megabytes; while the total size of the left-voices' web files is slightly over 1.13 Gigabytes. From the 1.875 Gigabytes of web files we were able to extract 399 Megabytes of political web log posts.

Since Gandelman's listing was dated March 2005, many of the web logs did not exist as far back as March 2003. Because of this the earlier datasets are in general smaller than the later dated datasets. Also, interest in our topic waxed and waned across the two-year period, affecting the sizes of the datasets.

3.2 Extracting Web Log Posts on Topic

We have chosen to limit the postings to a particular topic. It is the opinion of this topic we plan to identify. The topic we chose is people's opinion on how well President George W. Bush is handling the Iraq War. The topic of the posting is determined

by the terms: President Bush and Iraq War. Let t_1, t_2, t_3 ... t_n be the terms found within a posting p. The posting p is eligible for extraction if there exists t_i, t_j, t_k, t_l such that :

```
((t_i =~ "^President" || t_j =~ "^Bush") &&
(t_k =~ "^Iraq" || t_l =~ "^War")).                    (1)
```

The extraction rule is a perl regular expression that requires two concepts to be found within the extracted blog posting: President George W. Bush and the Iraq War. The rule allows either prefix terms *President* or *Bush* to represent the concept President George W. Bush. The Iraq War can be represented by prefix terms *Iraq* or *War*.

From the 399 Megabytes of web log posts, our topic selection filter determined 38,789 posts were deemed on-topic comprising 147 Megabytes, while 216,904 posts were deemed off-topic (252 Megabytes). As demonstrated by Table 1, the liberal bloggers consistently wrote more postings on-topic than the conservative bloggers; in some months the liberal posts outnumbered the conservative posts 2 to 1.

3.3 Dataset Representation

The datasets are represented by the most prevalent single word terms or *unigrams* occurring within the posts for the month. No stemming is performed on the terms. The features of the datasets are the unigrams occurring at least five times within the posting corpora. The values for the features represent presence versus absence of the feature within the post; we call this representation the *Boolean Presence feature set representation.* A value of 0 means the unigram was not found within the posting. Correspondingly, a value of 1 means the unigram was found within the posting. We chose the Boolean presence representation because it yielded a higher accuracy than the standard frequency feature representation in previous related research [14].

Since a unigram does not convey the context of a word, we used Das and Chen's technique to capture the polarity of the word's environment [3]. The idea is to negate words in the post that are found after a negative word such as *not*, or *no*. Since we are interested in sentiment, it is important we differentiate when words in a post are used to express the opposite meaning of the word. Unigrams are marked as negative if they are preceded by a negative term. The negative clause ends at the next punctuation mark. On average, this improves predictability between 2 to 4%.

We use a standard bag-of-features framework to represent our blog postings. Let $\{f_1,..., f_m\}$ be a predefined set of m features that may appear in a post. Let $f_i(d)$ be equal to 1 if the feature f_i appears in the post d and equal to 0 if the feature f_i does not appear in post d. Then each post d is represented by the post vector:

$$d = (f_1(d), f_2(d), ..., f_m(d)). \tag{2}$$

Table 1 lists the number of posts and the size of the feature sets for each month. The full feature set is created from all the posts within the month; while the reduced feature set is created from a randomly created category-balanced group of posts. The feature selection subset is determined by a forward feature selection algorithm that analyzes the utility of each feature. The selection algorithm seeks to remove redundant features.

Table 1. The percentage of postings on-topic, the number of postings, and the number of features for each month

Month	Percentage of Postings on Topic		Number of Postings		Number of Features		
	Right-voice	Left-voice	Right-voice	Left-voice	Full	Reduced	Feature Selection Subset
2003-03	16.92	24.48	258	400	9487	6150	133
2003-04	13.92	21.52	176	238	7059	4694	111
2003-05	9.46	18.02	113	208	6036	3501	105
2003-06	9.54	23.56	156	318	8023	4364	104
2003-07	11.60	29.79	207	464	9828	5264	120
2003-08	8.55	19.65	157	321	8792	4511	101
2003-09	13.45	25.71	257	424	10908	6149	141
2003-10	11.81	25.10	250	448	11028	6252	139
2003-11	15.62	25.33	276	410	10971	6602	142
2003-12	17.21	22.88	302	456	11884	6736	130
2004-01	14.30	23.21	352	636	13079	7574	157
2004-02	14.29	26.28	286	729	12318	6535	169
2004-03	15.43	26.48	370	819	13396	7994	123
2004-04	17.36	30.67	496	879	15729	9505	159
2004-05	15.41	29.98	440	1027	16767	9092	185
2004-06	16.95	27.32	417	902	16130	8697	219
2004-07	17.91	26.25	522	876	16565	9778	201
2004-08	17.84	27.62	615	1135	18819	11151	158
2004-09	21.53	32.75	784	1305	20644	12455	220
2004-10	23.55	31.83	972	1611	23009	14341	171
2004-11	15.95	20.84	467	807	17401	9408	197
2004-12	11.14	20.29	288	736	16254	6859	165
2005-01	13.77	22.12	506	971	18237	10245	209
2005-02	12.58	18.61	453	775	16285	9404	235
2005-03	12.26	15.69	336	633	14060	7822	149
Average	14.73	24.63	378	701	13708	7803	158

Table 1 provides some insights into the evolution of our topic over the two years. One striking statistic is the higher level of interest this topic has among the liberal bloggers than the conservative bloggers. Not only do we have more on-topic posts from the liberal bloggers, they also tend to post more often on this topic than the conservative bloggers. Also the number of posts on-topic varies from month to month. Some of this variation can be blamed on fewer blogs existing in March 2003 than in March 2005. However, the level of interest the liberal and conservative bloggers had in the current events of the war also accounted for the imbalance. On average, we had twice as many liberal posts as conservatives.

4 Machine Learning Techniques

We gauged the effectiveness of known sentiment classification technology on our novel collection of political web posts. We considered two different machine learning techniques: Naïve Bayes and Support Vector Machines and measured their applicability in our domain.

4.1 Naïve Bayes Classifier

A Naïve Bayes classifier is a probabilistic classifier based on probability models that incorporate strong independence assumptions among the features. Our Naïve Bayes classifier assigns a given web log post d the class c^*

$$c^* = \text{Argmax}_c\, P(c \mid d);\ c \in \{\text{right-voice, left-voice}\}. \tag{3}$$

A document of length n is represented as an m-dimensional vector, where f_i is the *ith* dimension in the vector and m is the number of features, as described in Section 3.3. We derive the Naïve Bayes (NB) classifier by first observing that by Bayes' rule

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \tag{4}$$

$P(d)$ plays no role in assigning c^*. To estimate the term $P(d \mid c)$, Naïve Bayes decomposes the estimate by assuming all the f_i's are conditionally independent given d's class. Term $n_i(d)$ is the presence of term i in document d (value 0 or 1).

$$P_{NB}(c \mid d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)})}{P(d)} \tag{5}$$

We chose to use a Naïve Bayes classifier because of its simplicity, its quick computation time compared to other machine learning techniques and its performance using the Boolean presence feature set representation in a previous study [14]. The Naïve Bayes assumption of attribute independence performs well for text categorization at the word feature level. When the number of features is large, the independence assumption allows for the parameters of each feature to be learned separately, greatly simplifying the learning process. The celerity of the Naïve Bayes modeling process makes it a favorable candidate for application to our fast-growing web log domain. Our experiments use the Naïve Bayes implementation from the WEKA machine-learning toolkit, version 3.4 [23]. We chose to use the Naïve Bayes' multinomial event-driven model.

4.2 Support Vector Machines

Support Vector Machines (SVMs) identify a hyperplane that separates two classes or categories of data. The chosen hyperplane creates the largest separation or margin between the two classes; hence it is a large margin classifier. Our search for the hyperplane is a constrained optimization problem. Assume we have n log posts to be

categorized. Our collection C of web log posts is represented as Formula 6 where x_i represents the features of the post; and c_i represents the categorization of that post, either a *left voice* or a *right voice*.

$$C = \{(x_1,c_1), (x_2,c_2),(x_3,c_3)...(x_n,c_n)\} \quad (6)$$

The dividing hyperplane of our two classes is defined to be w·x - b = 0. The parallel hyperplane for one category is defined as w·x - b = 1 and for the other category is w·x - b = -1. The space between the two parallel hyperplanes is the margin we wish to optimize. Not all of the data being classified is used in identifying the dividing hyperplane, only the closest points to the margin or the points that lie on the two parallel hyperplanes are used. These points are the contributing support vectors of the hyperplane. To include non-contributing points into the equations of the parallel hyperplanes, we rewrite the equations as inequalities, w·x - b >= 1 for one category and w·x -b <= -1 for the other category. The non-contributing data points will vary in distance from the corresponding hyperplane. Our two inequalities can be rewritten as Formula 7 since our c_i's represent the category values (1, -1) of our web posts. The quadratic optimization problem is to minimize the length of w given the constraint in Formula 7. This will identify the largest margin between our left and right voices.

$$c_i(\mathrm{w}\cdot\mathrm{x}_i - \mathrm{b}) >= 1 \quad \text{for } 1 < i < n. \quad (7)$$

We use the SMO kernel implementation from the WEKA machine-learning toolkit version 3.4 [23]. SMO, sequential minimal optimization, breaks the large quadratic optimization problem into the smallest quadratic optimization problems that can be solved analytically. We chose to use a SVM classifier because it outperformed other techniques in a previous study [14]. It also takes a different approach to classification than Naïve Bayes.

4.3 Validation Technique

We chose to use the same validation technique for all classifiers, stratified 10-fold cross-validation. In stratified 10-fold cross-validation, a dataset is randomly divided into 10 sets with approximately equal size and category distributions. For each *fold*, the classifier is trained using all but one of the 10 groups and then tested on the remaining group. This procedure is repeated for each of the 10 groups. The cross-validation score is the average performance across each of the ten runs.

4.4 Feature Selection

We investigated improving the collection of sentiment classifier's accuracy results by applying off-the-shelf feature selection to our datasets. In particular we have applied a forward search technique that evaluates the predictive ability of each feature individually and the redundancy among the features. The technique, *CfsSubsetEval* implemented in WEKA 3.4 [23], chooses a subset of the given features and aims to reduce the number of features while improving the accuracy results. We have chosen to search the feature set using a *BestFirst* search, starting from an empty subset and proceeding until the results of the current subset cannot be improved. The technique chooses features that are highly correlated with the predicting class but have low

intercorrelation. We chose this technique since we believe reducing redundancy within our features will support the Naïve Bayes assumption of independent features.

5 Experiments

In order to evaluate existing technology, we create seven different *collections of classifiers*, five containing Naïve Bayes classifiers and two containing Support Vector Machines. Each collection allows us to evaluate the effectiveness of one known aspect of the sentiment classification technology on our domain. Our goal is to achieve high accuracy on the results of the total dataset as well as on each of the two categories. We wish to keep our datasets small while still retaining high accuracies.

Our first collection of classifiers is created from *all* available posts from the left-voices and right-voices blogs. This collection contains datasets with different numbers of left-voices and right-voices log posts. We refer to it as our *unbalanced collection of classifiers*.

Our second collection of classifiers contains an equal number of left-voices and right-voices web log posts, but its feature set is determined by the full, unbalanced collection of datasets. We refer to it as our *balanced inflated collection of classifiers*. By comparing the results of our balanced inflated collection and our unbalanced collection, we can quantify the importance of balanced categories within our datasets.

Our third collection contains an equal number of left-voices and right-voices web log posts and its feature set is determined by this balanced dataset of posts. We refer to this collection as our *balanced collection of classifiers*. By comparing the results of our balanced collection to our balanced inflated collection, we can evaluate the two different feature set representations. It will reveal if more features on-topic improves the accuracy of the datasets.

Our fourth collection contains an unequal number of left-voices and right-voices posts. The categorical makeup is equivalent to the categorical makeup of the unbalanced collection of classifiers; however, the number of elements in each dataset is equivalent to the corresponding dataset in the balanced collection of classifiers. We refer to this collection as the *small unbalanced collection of classifiers*. We compare the results of these datasets to the results of the unbalanced collection of classifiers to consider the effects of unbalanced class constituency and dataset size to the accuracy results of left-voices and right-voices.

Our last Naïve Bayes collection contains an equal number of left-voices and right-voices posts. The feature set is determined by a subset feature selection technique described in Section 4.4. We refer to this collection as the *Naïve Bayes feature selection collection of classifiers*. We compare the results of this collection with the collection of balanced Naïve Bayes collection to consider the effects of our feature selection algorithm on our Naïve Bayes classifiers.

Our first Support Vector Machine collection contains an equal number of left-voices and right-voices posts, with the feature sets determined by the contributing posts and SVM classifiers. We refer to it as our *SVM collection of classifiers*. Comparing our balanced collection to our SVM collection of classifiers allows us to evaluate the effectiveness of our two machine learning techniques on our chosen domain.

Our next Support Vector Machines collection also contains an equal number of left-voices and right-voices posts; with the feature set determined by the *CfsSubsetEval* algorithm [23] described in Section 4.4. We refer to this collection as our *SVM feature selection collection of classifiers*. By comparing our SVM collection of classifiers to our SVM feature selection collection allows us to evaluate the effectiveness of our feature selection algorithm on our Support Vector Machine classifiers. We also compare our Naïve Bayes feature selection collection to the SVM feature selection collection to consider the effects of feature selection on our two chosen machine learning algorithms.

6 Results

Using our seven collections, Section 6.1 shows that Naïve Bayes performs well and SVMs perform adequately when predicting the sentiment of political blog posts even though the domain of our data is quite different from traditional text. In Section 6.2, we show increasing the feature set to contain topic-specific terms not selected by our feature selection algorithm does not improve the accuracy of the datasets; however decreasing the feature set to remove redundant features does improve the results of Naïve Bayes and Support Vector Machines. In particular on average it improves our Naïve Bayes results by 11.18% and our SVM results by 12.18%. We also show reducing the average size of the datasets by 30% in order to balance the categories does not have a negative effect on the total accuracy. It actually has the positive effect on the category makeup of the misclassified posts.

6.1 Comparing Different Machine Learning Techniques

Our first set of experiments compares two machine learning techniques: Naïve Bayes and Support Vector Machines on two collections of balanced datasets. In Figure 1, on average, SVMs correctly predicted the category of web log posts 75.47% of the time with a standard deviation of 2.64. Our Naïve Bayes classifiers outperformed Support Vector Machines, on average, by correctly predicting a posting's political category 78.06% of the time with a standard deviation of 2.39. We performed a paired samples t-test on our results, pairing our classifiers month-by-month. Our t-test showed Naïve Bayes outperforms SVMs at a 99.9% confidence level, with a confidence interval of [1.425, 3.488]. Previous research was able to achieve an 81.0% accuracy using Naïve Bayes and 82.9% using SVMs on a nonspecific corpus using the Boolean presence feature set representation [14]. SVMs are doing a poor job predicting the sentiment of our topic-specific web log posts compared to its success on a non-specific topic movie review corpus [14]. One potential cause for this is in our topic-specific corpus the number of terms in common between our two categories will be higher than in a nonspecific topic corpus. These common terms make it more difficult to identify the hyperplane separating the two categories; this finding contradicts Engström's results [5].

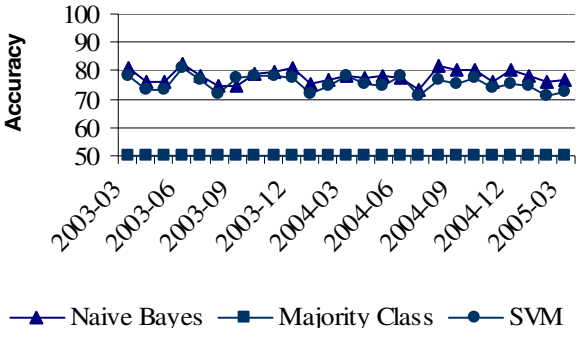

Fig. 1. Sentiment classification results of a collection of Naïve Bayes classifiers and SVM classifiers. Both sets contain the same data elements and feature sets.

6.2 Comparing Different Feature Sets

In Figure 2 we compare the collection of balanced classifiers to the collection of Naïve Bayes feature selection classifiers. In these sets of experiments the number of elements, the class composition, and the classifier, Naïve Bayes, remain constant. Only the feature set varies. As shown in Figure 2 the Naïve Bayes feature selection classifiers outperform the Naïve Bayes classifiers containing our baseline features. In Figure 3 we do the same comparison as in Figure 2, the only difference is the machine learning technique considered. We see improvement results in the SVM feature selection classifiers. In particular, our Naïve Bayes classifier collection coupled with a forward feature selection technique on average correctly predict a posting's sentiment 89.77% of the time with a standard deviation of 3.01. Our SVMs collection coupled with a forward feature selection technique on average correctly predicts a posting's sentiment 87.66% of the time with a standard deviation of 2.22. Naïve Bayes significantly outperforms Support Vector Machines at the 95% confidence level with a confidence interval of [1.5, 2.7]. On average, we gain an 11.84% increase for Naïve Bayes and a 12.18% increase for SVMs. These results show reducing the number of features by removing redundant features yields higher results for Naïve Bayes and SVM classifiers.

In Figure 4 we compare the collection of balanced classifiers to the collection of inflated balanced classifiers. In these sets of experiments the number of elements in the datasets is constant and the classifier is Naïve Bayes; only the number of features is varied. Our accuracy range for the collection of balanced inflated classifiers is 72.97% to 81.69%. The average predictability value is 78.06% with a standard deviation of 2.39. Our range for predictability for the collection of balanced classifiers is 73.16% to 82.67%, with an average predictability value of 77.93% and a standard deviation of 2.41. There is no improvement in accuracy with the inflated feature set even though the added features are relevant to the current month's data. The results for the two collections are indistinguishable. These results shows increasing the feature set with topic-related terms does not improve our results.

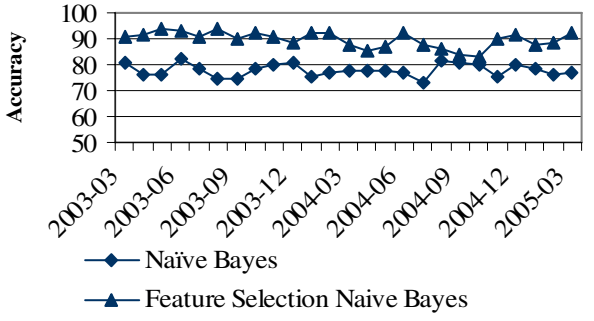

Fig. 2. Sentiment classification results of balanced Naïve Bayes classifiers vs. Feature Selection Naïve Bayes. The feature selection Naïve Bayes classifiers significantly outperform the Naïve Bayes classifiers.

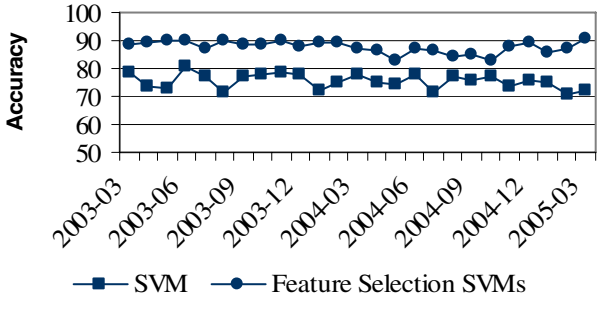

Fig. 3. Sentiment classification results of balanced Support Vector Machine classifiers vs. Feature Selection SVMs. The feature selection SVM classifiers significantly outperform the SVM classifiers.

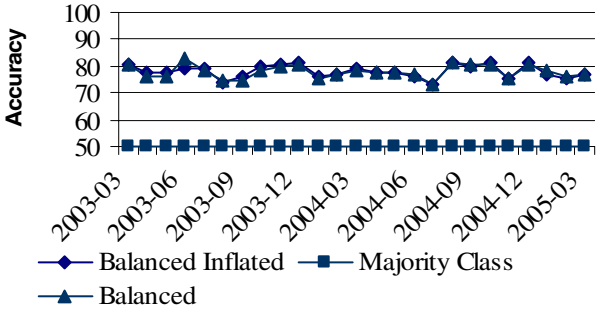

Fig. 4. Sentiment classification results of two sets of balanced Naïve Bayes classifiers vs. the Majority class. The difference between the two balanced sets is the number of features used.

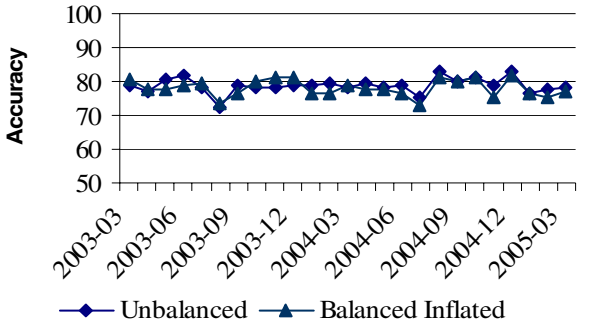

Fig. 5. Sentiment classification results of a set of balanced inflated classifiers and a set of unbalanced classifiers. The sets have identical feature sets.

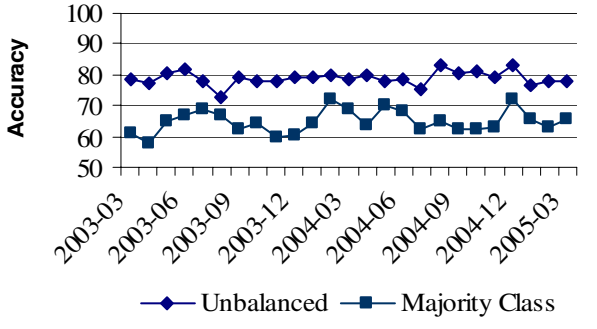

Fig. 6. Sentiment Classification results of a set of unbalanced Naïve Bayes classifiers compared to the actual percentage of the dataset belonging to the Majority Class

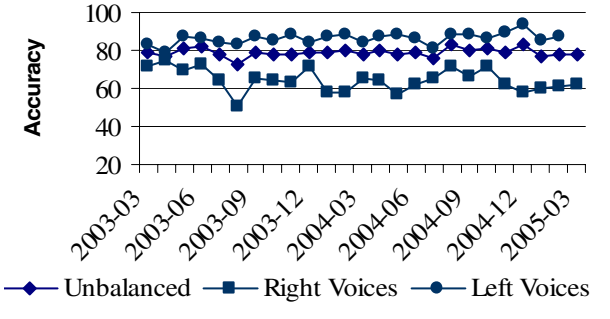

Fig. 7. Sentiment Classification results by category of a set of unbalanced Naïve Bayes Classifiers by category

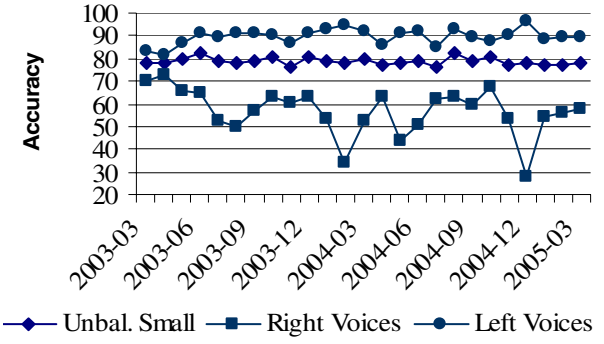

Fig. 8. Sentiment classification results by category of a set of smaller, unbalanced Naïve Bayes classifiers. Note the change in range of the y axis from the above graphs.

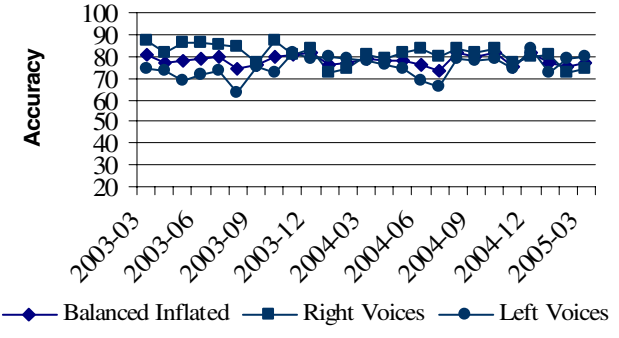

Fig. 9. Sentiment classification results by category of a set of balanced inflated Naïve Bayes classifiers. To ease comparison to Figure 8, this graph has an extended y axis range.

6.3 Comparing Different Categorical Constituencies

Figure 5 compares the results of the balanced inflated classifiers to the unbalanced classifier results. In these sets of experiments the collections contain Naïve Bayes classifiers with identical feature sets. Our unbalanced collection of classifiers contains all on-topic log posts that were available for the given months. Even though the sizes of the balanced datasets are on average only 70% the size of the corresponding unbalanced datasets, Figure 5 illustrates that the total accuracy of the two sets are strikingly similar; they are within fractions of each other.

Yet Figure 6 shows the unbalanced classifiers in many months are barely outperforming the Majority class found within the datasets. We wanted to explain the poor results from our unbalanced classifiers. We believe the answer lies in the constituency of the correctly classified instances rather than in a category that is intrinsically more

difficult to predict. To understand this observed effect, we begin by comparing the success in predictability of the two categories (left-voices and right-voices) to the accuracy of the total population as shown in Figure 7. Clearly, we are doing a poor job on our right-voices; our category containing fewer posts. The left-voice category consistently outperforms the right-voice category.

The discrepancy in predictability between our two categories can be attributed to the imbalance in our datasets, as we can show by the following two sets of experiments. The first set of experiments keeps the constituency of the datasets constant and varies the size of the datasets. Our next set of experiments varies the constituency of the datasets while keeping the dataset size constant. Both sets of experiments contain the same Naïve Bayes induction technique and the same feature set representation.

The results of our same class constituency and smaller dataset size experiments are displayed in Figure 8. These smaller datasets performed worse on predicting the right-voice postings than the original unbalanced classifiers. The average accuracy for the right-voice category in the larger unbalanced dataset was 64.34%, for the smaller unbalanced dataset 56.71%. The average accuracy for our left-voice category in the collection of larger unbalanced classifiers was 86.30%, for the smaller unbalanced dataset 89.58%. As the dataset size decreases the effect of the imbalanced class makeup of the datasets dramatically increases the bias found within the correctly classified posts.

In Figure 9, we vary the constituency of the datasets, while keeping the size constant. As shown in the figure, in some months the left-voices are easier to predict while in other months the right-voices are predicted more accurately. The overall average for the left-voices category is 75.09% for the right-voices category is 80.82%. We generated the overall average of the individual month's percentage of misclassifications per category; the left-voice category constitutes 56% of the misclassified posts while the right-voice category constitutes 44%. When given a uniform distribution in the datasets, right-voices are easier to predict than left-voices. This is especially true for the early segment of the time spectrum, or the first months of the war from March 2003 to November 2003. In this section the left-voice category constitutes 64% of the misclassified posts while the right-voice category constitutes 36%.

Figure 8 and 9 together demonstrate reducing the average size of the datasets by 30% in order to balance the categories did not have a negative effect on the total accuracy. It actually had the positive effect on the category makeup of the misclassified posts.

7 Conclusions and Future Work

We have investigated the utility of Naïve Bayes and SVMs on a novel collection of datasets created from political web log posts. We showed a Naïve Bayes classifier significantly outperforms Support Vector Machines at a confident level of 99%, with a confidence interval of [1.425, 3.488]. We show applying feature selection to our results can improve our results significantly, in particular it improves our Naïve Bayes

results by 11.84% and our SVM results by 12.18%. We show a Naive Bayes classifier is sensitive to the class makeup of the dataset. Not having a balanced composition of the classes introduces a bias within the results; the majority class is more likely to be classified correctly than the minority class. As the databases decrease in size, the bias effect due to the unbalanced composition of the datasets magnifies.

We also showed our baseline feature set representation works as well as a similar feature set representation that was on average 1.75 times larger than our representation. The larger feature set was generated from all the on-topic web log posts for the current month. The added features were all from left-voices web posts. However, the added features did not improve the accuracy of the classification of the left-voices posts.

We have shown we can predict the political leanings of a posting on the Iraq War at an average accuracy of 78.6% for a two-year period without feature selection technique and 89.77% on average with a forward search feature selection technique. Even though we have not tried another topic we believe we would attain similar results on another topic since there is nothing particular in our sentiment classification system approach that is particular to our chosen topic.

There are many interesting questions we can explore with our current dataset, including different time partitions, different representations for our postings, different representations for the feature sets, and different values for those features. We can explore the effects of size posting on predictability. Finally, we would like to further our research by exploring the ability to track changes within people's opinions on a particular topic and explore the time dependency of our data. We want to be able to classify the data within months as stable (consistent with previous data), or trendy (not pertaining to previous discussions). We are also interested in identifying the length of trends within the data.

Acknowledgements. We thank Stuart M. Shieber for his insightful comments on this work. This research was supported in part by a research gift from Google.

References

[1] Beineke, P., Hastie, T., Vaithyanathan, S.: The Sentimental Factor: Improving Review Classification via Human-Provided Information. In: ACL 2004. Proceedings ACL: Association of Computational Linguistics, Barcelona, pp. 263-270 (2004)

[2] Carroll, J.: Local TV and Newspapers Remain Most Popular News Sources, Increased use of Internet news this year. The Gallup Poll. poll.gallup.comcontent/default.aspx?CI=14389 (December 2004)

[3] Das, S., Chen, M.: Yahoo! for Amazon: Extracting Marketing Sentiment from Stock Message Boards. In: APFA 2001. Proceedings of the 8th Asia Pacific Finance Association Annual Conference (2001)

[4] Dube, J.: Blog Readership up 58% in 2004. CyberJournalist.net (January 2005), www.cyberjournalist.net/news/001819.php

[5] Engström, C.: Topic Dependence in sentiment classification. Master's thesis, St Edmunds's College, University of Cambridge (2004)

[6] Gard, L.: The Business of Blogging. Business Week Online (December 2004)

[7] Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of the ACL-EACL 1997 Joint Conference: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, pp. 174–181 (1997)
[8] Hearst, M.: Direction-based text interpretation as an information access refinement. In: Jacobs, P. (ed.) Text-Based Intelligent Systems, Lawrence Erlbaum Associated (1992)
[9] Hu, M., Liu,B.: Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining KDD 2004, pp.168-174 (2004)
[10] Huettner, A., Subasic, P.: Fuzzy typing for document management. In: ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pp. 26–27 (2000)
[11] Kushal, D., Lawrence, S., Pennock, D.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: WW W 2003. Proceedings of the Twelfth International World Wide Conferences, pp. 519–553 (2003)
[12] Madden, M.: Online Pursuits: The Changing Picture of Who's Online and What They Do. Pew Internet and the American Life Project Report (2003), www.pewinternet.org/PPF/r/106/report_display.asp
[13] Nasukawa, T., Yi, J.: Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In: Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture, pp. 70–77 (2003)
[14] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86 (2002)
[15] Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of the 42nd ACL, pp. 271-278 (2004)
[16] Pew Internet and the American Life Project (2004), www.pewinternet.org/trends/Internet%20Activities_12.21.04.htm
[17] Pew Internet and the American Life Project (2005), www.pewinternet.org/trends/Internet_Activities_12.05.05.htm
[18] Rainie, L.: The State of Blogging. Pew Intenet and the American Life Project Report (2005), www.pewinternet.org/PPF/r/144/report_display.asp
[19] Rainie, L., Shermak J.: Search engine use shoots up in the past year and edges towards email as the primary internet application. Pew Internet and the American Life Project Report in conjunction with comScore Media Metrix (2005), www.pewinternet.org/pdfs/PIP_SearchData_1105.pdf
[20] Sack, W.: On the computation of point of view. In: Proceedings of the Twelfth American Association of Artificial Intelligence (AAAI), pp. 1488. Student Abstract (1994), www.pewinternet.org/pdfs/PIP_SearchData_1105.pdfv
[21] Tong, R M.: An Operational System for Detecting and Tracking Opinions in On-line Discussion. In: SIGIR 2001 Workshop on Operational Text Classification (2001)
[22] Turney, P.D., Littman, M.L.: Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus. Technical Report EGB-1094, National Research Council Canada (2002)
[23] Witten, I.H., Frank, E.: Data Mining Practical Learning Tools and Techniques with Java Implementations. Academic Press, San Diego, CA (2000)

Appendix A: Web Logs Used in this Research

Liberal Web Logs	Conservative Web Logs
aboutpolitics.blogspot.com	atrainwreckinmax-
allspinzone.blogspot.com	well.blogspot.com
www.americablog.org	acepilots.com
www.reachm.com/amstreet	www.alarmingnews.com
angrybear.blogspot.com	www.alittlemoretotheright.com
atrios.blogspot.com	alwaysright.blogs.com
www.bopnews.com	always_right
www.bullmooseblog.com	americandigest.org
www.burntorangereport.com	anticipatoryretaliation
www.busybusybusy.com	armiesofliberation.com
cernigsnewshog.blogspot.com	asmallvictory.net
corrente.blogspot.com	www.balloon-juice.com
www.crookedtimber.org	betsyspage.blogspot.com
www.cursor.org	www.blogsforbush.com
www.dailykos.com	www.blogsofwar.com
www.davidsirota.com	www.bobhayes.net
demagogue.blogspot.com	bogusgold.blogspot.com
www.democraticunderground.com	www.calblog.com
demwatch.blogspot.com	coldfury.com
digbysblog.blogspot.com	command-post.org
dneiwert.blogspot.com	commonsense-
emergingdemocraticmajoritywe-	runswild.typepad.com
blog.com	www.littlegreenfootballs.com
donkeyrising/index.php	mypetjawa.mu.nu
donkeywonk.blogs.com/mrleft	northeastdilemma.blogspot.com
nielsenhayden.com/electrolite	pikespeak.blogspot.com
ezraklein.typepad.com/blog	www.thepoliticalteen.net
farleft.blogspot.com	talesofawanderingmind
geffen.blogspot.com	www.slantpoint.com
	www.slingsnarrows.com/blog
www.heartsoulandhumor.blogspot.com	www.qoae.net
www.hoffmania.com	www.redlinerants.com
jackotoole.net	redmindbluestate.blogspot.com
jameswolcott.com	rightmoment.blogspot.com
www.joeterrito.com	www.right-thinking.com
www.juancole.com	rightwingnews.com
kbonline.typepad.com/random	sayanythingblog.com
kirghizlight.blogspot.com	www.sgtstryker.com
www.kudzufiles.com	www.shotinthedark.info
lastonespeaks.blogspot.com	southernappeal.blogspot.com
www.leanleft.com	principledobjection.blogspot.com
www.liberaloasis.com	www.thewaterglass.net
www.liquidlist.com	varifrank.com
markschmitt.typepad.com	volokh.com
maxspeak.org/mt	wizbangblog.com
mediamatters.org	xrlq.com
www.michaeltotten.com	youngpundits.blogspot.com
moderateleft.blogspot.com	themarylandmoder-
www.mydd.com	ate.blogspot.com

www.nathannewman.org/log	therapysessions.blogspot.com
newleftblogs.blogspot.com	www.danieldrezner.com/blog
wpblog.ohpinion.com	www.davidlimbaugh.com
www.oliverwillis.com	demrealists.blogspot.com
www.pandagon.net	www.diggersrealm.com
www.patridiots.com	www.donaldsensing.com
www.pennywit.com/drupal/index.php	www.eddriscoll.com/weblog.php
presidentboxer.blogspot.com	www.erickerickson.org
profgoose.blogspot.com	www.fringeblog.com
www.prospect.org/weblog	www.gaypatriot.org
www.richardsilverstein.com	www.hughhewitt.com
rittenhouse.blogspot.com	www.hundredpercenter.com
rogerailes.blogspot.com	incite1.blogspot.com
rogerlsimon.com	www.indcjournal.com
roxanne.typepad.com/rantrave	www.indepundit.com
samueljohn-	www.instapundit.com
son.com/blog/otherblog.html	www.inthebullpen.com
seetheforest.blogspot.com	www.iraqnow.blogspot.com
stevegilliard.blogspot.com	www.jquinton.com
suburbanguerrilla.blogspot.com	justoneminute.typepad.com
www.tbtmradio.com/geeklog	lashawnbarber.com/index.php
www.talkingpointsmemo.com	libertariangirl.blogspot.com
www.talkleft.com	www.gregpiper.com
tbogg.blogspot.com	conservativeeyes.blogspot.com
thatcoloredfellasweblog.bloghorn.com	www.dailynewsbrief.com
www.the-hamster.com	dailypundit.com
www.theleftcoaster.com	www.danegerus.com/weblog
www.thetalentshow.org	www.calicocat.com
www.thetalkingdog.com	cbcbcbcb.blogspot.com
thinkprogress.org	chrenkoff.blogspot.com
www.thismodernworld.com	
www.tompaine.com/blogs	
www.unspun.us	
usliberals.about.com	
wampum.wabanaki.net	
warandpiece.com	
www.washingtonmonthly.com	
xnerg.blogspot.com	