

Finding Political Blogs and Their Political Leanings

Maojin Jiang
Linguistic Cognition Lab
Illinois Institute of Technology
Chicago, IL 60616 USA
jianmao@iit.edu

Shlomo Argamon
Linguistic Cognition Lab
Illinois Institute of Technology
Chicago, IL 60616 USA
argamon@iit.edu

Abstract

The Blogosphere has more influence on both the general public's opinions and mainstream media nowadays. As vast political blogs representing grassroots voices go online, it is both interesting and useful to find them and determine their political leanings as either liberal or conservative. In this paper, we address both problems by using front pages of blogs in a corpus we built to create two classifiers: one to classify arbitrary blogs as political or non-political; the other to classify political blogs as liberal or conservative. We explored performance by using 'bag of words' and link features with the help of the SVM classification algorithm. Results show that a political vs. non-political blog classifier can achieve an overall accuracy of 91.40% by merely using 'bag of words' features. Combining both 'bag of words' and some link features, the best liberal vs. conservative blog classifier can on average predict a political blog's political leaning with an accuracy of 86.63%. With the help of some feature selection methods, however, analysis also indicates that potential in using simple 'bag of words' features in political leaning categorization is limited in that it lacks enough discerning power in detecting differences of opinions expressed by liberal bloggers from conservative ones.

1 Introduction

Simply put, a blog is a personal, business or government journal or e-diary website for public access, usually displayed in reverse chronological order. With the popularity of the Internet, the Blogosphere has increasing influence on the general public's opinions and mainstream media (MSM) nowadays. According to blogherald, as of July 2005, there were over 70 million blogs, 90% of which were created on eight leading blog hosting sites that hosted one million or more blogs alone. In the United States, the number of blogs is approximately 15-30 million, which means that one out of every 10 people owns a blog (according to 2000 census, the population of the U.S. is nearly 300,000,000).

Thanks to the low cost of blogging and the lack of geographical barriers for developing and maintaining social networks, more and more bloggers have already carried out the roles of grassroots reporters and fact-checkers [15]. In this regard, political blogs have also taken on increasingly important roles in our daily lives. In addition, sometimes political blogs can carry a special

mission that no other type of blogs can have. According to a survey project conducted by the Institute for Politics, Democracy & the Internet, "political blogs have also been influential in raising money for political candidates and pushing select races into the national spotlight" [16].

As a result, it is no surprise that some 2008 Presidential candidates have included blogs on their websites, and that there are even more blogs on the Internet maintained by their supporters.

Unfortunately, among all political blogs, only a small number of blogs can catch readers' eyes. Those daily political blog readers only visit the most popular blogs according to [16]. However, many more political blogs represent grassroots voices. On one hand, there exists a desire from ordinary bloggers that their voices will be widely heard; on the other hand, governments want to know the thoughts of their citizens. Thus, it becomes an important task to find political blogs buried in the Internet 'forest' and locate bloggers' positions along the political leaning continuum, from liberal to conservative.

Though text categorization has been well researched, previous work on political blog categorization has been very limited. As a result, it has motivated us to study this problem due to its important implication in the real world. In this paper, we will describe our approaches to this problem and present some preliminary results.

2 Previous work

2.1 Political text analysis A lot of work has been seen in analyzing political text. For example, [37] investigated transcripts of U.S. Congressional floor debates to determine whether the speeches represented support of or opposition to proposed legislation. [36] analyzed large collections of text commentary to help the United States federal government's regulation writers to formulate the rules and regulations that define the details of laws enacted by Congress. [25] presented

ways to extract policy positions from political texts. In [1], authors analyzed citation patterns by examining intra- and inter-community links and studied some textual content similarities of liberal and conservative blogs over a period of two months preceding the U.S. Presidential Election of 2004.

Many other research activities in this area have been conducted in the form of *sentiment analysis*, which aims to find, extract and categorize opinions represented in political texts. In [33, 28], authors tried to classify political orientations in online political discussion postings. Similarly, [9] performed sentiment classification on blog postings relevant to certain topics.

2.2 Blog categorization Blog (or Blogger) categorization has become a more popular area in web mining recently. One important application problem is to find online communities by categorizing similar blogs or bloggers into groups in order to perform social network analysis [14, 13, 18, 8, 41, 19, 26]. [6, 22] explored different methods in text categorization of blogs in terms of similarity of topics in general. [2, 24] analyzed significant differences in writing style and content between male and female bloggers as well as among authors of different ages, which lends tools for blog categorization based on a blogger’s age and gender.

Another popular blog categorization is sentiment analysis. [31, 3] identified bloggers’ moods by capturing salient linguistic features with the help of machine learning algorithms. In [33, 28, 9], authors aimed to classify certain blog postings into categories of opposite political orientations.

Lastly, as a result of emerging blog spams, blog spam detection and filtering has also become popular recently [34, 17, 30, 23].

3 Our approach

Our approach to political blog categorization is to use state-of-the-art machine learning algorithms to learn two classifiers that classify blogs into different categories. The first classifies arbitrary blogs as political or non-political blogs, and the second determines political leanings of political blogs and classifies them as liberal or conservative. The purpose of building the first classifier is to identify political blogs for applying the second classifier. That is, given an arbitrary blog, we first determine if it is a candidate of liberal or conservative blog by applying the first classifier. If the answer is positive, we will then further determine its political leaning by using the second classifier.

As our first investigation into the problem, our goal was to explore preliminary performance by using two basic features and their combinations, including bag of

words (BoW) and out-links of *front pages* of blogs. The reason that only *front pages* are used is that they reflect the latest postings of bloggers and are thus “current”. If the whole website of a blog is used, too much irrelevant information will be introduced as “noise”. To achieve this goal, we created a blog corpus for training and testing. In the following subsections, we will describe our approach in more detail.

3.1 Corpus Our corpus consists of three groups of blogs, namely, liberal blogs, conservative blogs and non-political blogs. We obtained addresses of these blogs from five blog catalog websites. On different catalog websites, different category naming conventions are used. For our purpose, liberal blogs in our corpus include those labeled as “left”, “liberal”, “liberalism” and “democratic”; conservative blogs include those labeled as “right”, “conservative”, “conservatism” and “republican”. In addition, we downloaded blogs in 41 different categories that we believed were least relevant to politics.

Originally, from these five blog catalog websites, there were 1428 liberal blogs, 1099 conservative blogs and 2809 non-political blogs. The distribution of these liberal and conservative blogs is shown in Table 1, in which the first column shows URLs of blog catalog websites, the second column indicates category names used by blog catalog websites, and the third column gives the total number of blogs under the category shown in the second column. Table 2 lists all 41 non-political categories.

Catalog website	Label	#Blogs
blogcatalog.com	Democratic	818
	Republican	666
blogarama.com	Left	473
	Right	303
etalkinghead.com	Liberal	133
	Conservative	126
campaignsandelections.com	Liberal	63
	Conservative	48
blogs.botw.org	Liberalism	78
	Conservatism	67

Table 1: Liberal and Conservative blogs originally listed at five blog directory websites

In order to obtain a high-quality corpus for training and testing, after *front pages* of original 1428 liberal blogs and 1099 conservative blogs were downloaded, we manually examined the validity of each to remove “noise” blogs. After this process, we identified and removed the following “noises”: 1) non-English blogs, 2) non-existent, domain expired or relocated blogs, 3)

Academics	Computers	Gaming	Outdoor	Personal	Travel	Environment	Music
Art_Artists	Technology	Health	Lifestyle	Pets	Vacation	Shopping	Celebrity
Arts_Entertainment	Crafts	History	Loans	Philosophy	Writing	Finance	Investing
Education_Training	Autos	Home_Garden	Meme	Places	Business	News_Media	Sports
Entrepreneurship	Humor	Mobile	Science	Career_Jobs	Internet	Coaching	Food_Drink
Law_Legal							

Table 2: 41 non-political blog categories

short blogs less than 140 words (excluding HTML tags), 4) portal blogs, which only have listing of categories without any postings, and 5) non-political blogs (it turns out that these are mostly commercial web sites disguised as either liberal or conservative blogs). As a result, 374 ‘liberal’ blogs and 306 ‘conservative’ blogs were removed in total (including duplicate blogs and those labeled as both “liberal” and “conservative”), although we did not discard any non-political blogs. In the end, our corpus contained *front pages* of 1054 liberal blogs, 793 conservative blogs and 2809 non-political blogs, totalling more than 200MB.

At this point, we have to admit that in the real world, an arbitrary blog to be categorized may be one type of these discarded “noise” blogs. However, these types of blogs are excluded from the corpus for evaluation, therefore, one might have reason to doubt any result obtained against the corpus by using the methods developed in this paper. We noticed this issue and provided the means to automatically remove certain “noises”. We will address this in section 4.3.

Finally, we assigned a “popularity” value to each liberal and each conservative blog and ranked liberal and conservative blogs separately according to their popularity values. The popularity value of a blog comes from two measures. One is *traffic* of the blog, calculated as the number of visits per day, tracked by ‘The Truth Laid Bear’¹. The other is *citation* of the blog, calculated as the number of pages that link to the blog, retrieved from Google. Between these two measures, *traffic* is less reliable and covers a fewer number of blogs because it only tracks those registered with ‘The Truth Laid Bear’. It is thus expected that some blogs in our corpus would not appear on the *traffic* list. Instead, *citation* is more trustworthy. Thus, we linearly combined the two measures into a popularity value for each blog, giving more weight to *citation*. Given a blog θ of a political leaning x , let $MAX_traffic(x)$ be the maximum traffic of all blogs with political leaning x in our corpus, let $MAX_citation(x)$ denote the maximum citation of all blogs with political leaning x in our corpus, the popularity of θ_x , $pop(\theta_x)$ is calculated as follows:

$$pop(\theta_x) = 0.2 * \frac{traffic(\theta_x)}{MAX_traffic(x)} + 0.8 * \frac{citation(\theta_x)}{MAX_citation(x)}$$

Table 3 shows top 10 liberal blogs with traffic and citation values and Table 4 shows top 10 conservative blogs.

Rank	Blog	Citation	Traffic
1	www.dailykos.com	47900	458497
2	atrios.blogspot.com	41000	80321
3	www.crooksandliars.com	35000	158849
4	www.washingtonmonthly.com	37600	37385
5	www.juancole.com	32200	0
6	crookedtimber.org	30000	0
7	www.pandagon.net	27100	11210
8	digbysblog.blogspot.com	26100	24069
9	www.talkleft.com	26200	9992
10	www.firedoglake.com	23600	47413

Table 3: Top 10 Liberal blogs

Rank	Blog	Citation	Traffic
1	www.instapundit.com	64500	220445
2	www.michellemalkin.com	37400	145827
3	www.littlegreenfootballs.com	28200	103836
4	www.powerlineblog.com	34700	53608
5	www.redstate.com	20600	29871
6	www.balloon-juice.com	12700	20107
7	www.conservativenation.us	0	23665
8	www.rightwingnews.com	19700	10792
9	patterico.com	12900	8564
10	hughewitt.townhall.com	28300	0

Table 4: Top 10 Conservative blogs

3.2 Feature sets In our current work, we used two feature sets, BoW and out-links, described as follows, to represent a blog.

BoW: Full “Bag of Words” feature set, with one binary feature for each different word type; a feature value is 1 if the corresponding word occurs in the blog, and 0 otherwise. In our experiments, there were 398,933 such features.

¹<http://truthlaidbear.com>

Out-link: As we mentioned earlier, all liberal blogs and all conservative blogs are ranked according to their “popularity”. From the ranked list of liberal blogs, we created four out-link features: ‘Ltop10’, ‘Ltop20’, ‘Ltop30’ and ‘L’. Each of them takes on a binary value: 1 or 0. Specifically, if a blog contains at least a link to a liberal blog in the corpus that is ranked among top 10, or ranked among top 11 to top 20, or ranked among top 21 to top 30, or ranked beyond top 30, then ‘Ltop10’, or ‘Ltop20’ or ‘Ltop30’, or ‘L’ takes on a 1 respectively. Otherwise, they all take on 0. However, a link to a blog itself is excluded from generating out-link features for this blog. Similarly, four out-link features are created from ranked list of conservative blogs, namely, ‘Rtop10’, ‘Rtop20’, ‘Rtop30’ and ‘R’. So, in total, there are eight binary out-link features.

3.3 Classifier learning As we mentioned earlier, we modeled the blog categorization problem as creating two classifiers. One is to classify arbitrary blogs into two categories: “political” blogs (more accurately speaking, candidates are of either liberal or conservative blogs), and “non-political” blogs. The other is to classify “political” blogs as either ‘liberal’ or ‘conservative’. Toward this goal, blogs are processed and represented as a vector of values of the features selected. Certain blogs represented in this way are used as a training set. Then, building classifiers are conducted by using standard machine learning techniques that have been used successfully for text classification tasks in the past. We report here on results using classification in Joachims’s *SVM^{light}* implementation of SVM [21]. This tool can handle problems with thousands of support vectors efficiently, has scalable memory requirements, and has many useful options for learning.

3.4 Feature selection It is our intuition that liberal or conservative bloggers are often concerned with different things, and they tend to hold different opinions on common issues. As a result, these differences are reflected in their blogs and demonstrated at least by their word usage. The same phenomenon can be observed when looking at political blogs and non-political ones.

In order to investigate this assumption, we used five feature selection methods to identify *discriminating words* that distinguish one category of blogs from another, that is liberal vs. conservative, and political vs. non-political. We only applied feature selection methods to BoW features. In addition to the aforementioned requirement from the application point of view, another benefit of using feature selection is dimensionality re-

duction, which can reduce computing cost and enhance classification performance by avoiding overfitting.

The first feature selection method is *F-score*². In [5], authors combined it with other feature selection methods in their SVM tools and were able to recommend a feature size to be used. The other four methods are based on statistics and information theory, including: *chi-square* (CH), *odds ratio* (OR), *information gain* (IG), and *mutual information* (MI) [35]. The functions of these feature selection methods are listed in Table 5. In the table, t_k denotes a BoW feature; C_i is a binary category, either C_1 or C_2 ; $\mu(t_k, C_1)$ and $\mu(t_k, C_2)$ represents average of feature t_k in category C_1 and C_2 respectively; $\mu(t_k)$ is average of feature t_k in both C_1 and C_2 . $(t_{k,j}, C_1)$ and $(t_{k,j}, C_2)$ are the value of feature t_k in the j th instance in C_1 and C_2 respectively. Other notations follow the same meanings in [35].

From the table, it can be seen that *F-score* is a ‘global’ function in that it assigns a value to a feature t_k by considering all categories; instead, all other four methods assign each feature a value ‘locally’ to a category C_i . To obtain a ‘global’ discriminating power for each feature with these methods, ‘local’ values should be aggregated. There are two aggregation methods to determine this ‘global’ value : 1) ‘sum’, summation of ‘local’ values; and 2) ‘max’, maximum ‘local’ value.

After each feature is assigned a ‘global’ value, all features can be ranked according to these values. A certain number of top features can then be selected in building classification models based on expected performances and are deemed as “discriminating words”. To determine the number of features to be selected, one can start building a model with all features, then iteratively reduce the feature size in half by choosing the top half features if performance improves or does not drop beyond a threshold.

3.5 Evaluation methods Traditional metrics in text categorization are used in evaluating both classifiers. First, we calculate *precision*, *recall* and *F-score* for each category. Precision is calculated as the number of correctly classified blogs in a category divided by the total number of blogs classified under that category. Recall is calculated as the number of correctly classified blogs in a category divided by the total number of blogs actually belonging to that category. In our binary blog categorization, each category is equally important, recall of classifying blogs into one category directly impacts the precision of classifying blogs into the other and vice versa. So, we calculate *F1* measure as *F-score*

²<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#3>

Selection method	Denotation	Function
F-score	$F(t_k)$	$\frac{(\mu(t_k, C_1) - \mu(t_k))^2 + (\mu(t_k, C_2) - \mu(t_k))^2}{\frac{1}{ C_1 - 1} \sum_{j=1}^{ C_1 } ((t_{k,j}, C_1) - \mu(t_k, C_1))^2 + \frac{1}{ C_2 - 1} \sum_{j=1}^{ C_2 } ((t_{k,j}, C_2) - \mu(t_k, C_2))^2}$
Chi-Square	$\chi^2(t_k, C_i)$	$\frac{ T_r \cdot [P(t_k, C_i)P(\bar{t}_k, \bar{C}_i) - P(t_k, \bar{C}_i)P(\bar{t}_k, C_i)]^2}{P(t_k)P(\bar{t}_k)P(C_i)P(\bar{C}_i)}$
Odds Ratio	$OR(t_k, C_i)$	$\frac{P(t_k C_i)(1 - P(t_k \bar{C}_i))}{(1 - P(t_k C_i))P(t_k \bar{C}_i)}$
Information Gain	$IG(t_k, C_i)$	$P(t_k, C_i) \log \frac{P(t_k, C_i)}{P(C_i)P(t_k)} + P(\bar{t}_k, \bar{C}_i) \log \frac{P(\bar{t}_k, \bar{C}_i)}{P(\bar{C}_i)P(\bar{t}_k)}$
Mutual Information	$MI(t_k, C_i)$	$\log \frac{P(t_k, C_i)}{P(t_k)P(C_i)}$

Table 5: Feature selection function

for each category, as follows:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Second, overall accuracy is calculated to evaluate how many blogs of both categories in total are correctly classified. It is calculated as the number of correctly classified blogs of both categories divided by the total number of blogs to be classified. The average $F1$ measure is another way to evaluate overall performance. Since both categories are of equal importance, the average $F1$ measure is simply calculated as the average of summation of $F1$ of each category classification.

4 Experimental evaluation

Two sets of experiments have been performed: one for building and testing models to classify blogs as political vs. non-political and the other for liberal vs. conservative blog classifier. We first built each classifier by using all BoW features. Then, we applied each feature selection method to find out the discriminating words and improve on both effectiveness and efficiency of the final models. Next, we added out-link features to see how they influenced results, after which we discussed imbalanced data set problems and presented our methods with experimental results. Lastly, at the end of this section, we proposed some methods and conducted experiments on detecting and discarding “noise” blogs, which cope with the issue mentioned in the section of describing corpus (section 3.1). All results reported in this paper were obtained by using SVM^{light} with linear kernel in a 10-fold cross-validation (using a more complicated kernel did not achieve any significant improvement, as a result, it is reasonable to use the simplest method). Since categorizing blogs in terms of political leanings is of greater interest to us, we will present the results of political leaning classification first, followed by political blog classification.

4.1 Political leaning classification

4.1.1 Using all BoW features Firstly, we evaluated the results solely by using all BoW features. In this experiment, all 1054 liberal blogs were used to generate positive class instances; all 793 conservative blogs were used to generate negative class instances. They were combined in a 10-fold cross-validation. Table 6 shows the results.

Category	Precision	Recall	F1
Liberal	73.42%	94.13%	0.8250
Conservative	87.71%	54.94%	0.6757
Average F1:			0.7503
Overall accuracy:			77.20%

Table 6: Liberal vs. Conservative classification (all BoW features)

4.1.2 Feature selection Next, we applied all feature selection methods to the top 100 liberal blogs and the top 100 conservative blogs as positive and negative classes. Only the remaining 954 liberal blogs and 693 conservative blogs were used in 10-fold cross-validation experiments. We used both ‘sum’ and ‘max’ in aggregating ‘local’ values for each feature in all feature selection methods except F -score. These results are comparable by using either aggregation method. So, we only show results by using ‘sum’ for chi -square, odds ratio, information gain and mutual information in Table 7 together with F -score, indicating the optimal feature size in terms of achieving the highest ‘average F1’. It can be seen that all these methods except $mutual$ information generated comparable results.

Comparing Table 7 to Table 6, we can see that all feature selection methods are helpful except $mutual$ information. This seems to be consistent with previous work in text categorization. For example, [39] reported that $information$ gain with ‘sum’ aggregation method and chi -square with ‘max’ aggregation method can reduce feature dimensionality by a factor of 100 with no loss of effectiveness (sometimes even with a slight

Feature selection	Feature size	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	\bar{F}_1
Information Gain (IG)	6963	75.95%	91.61%	0.8305	84.31%	60.53%	0.7047	78.42%	0.7676
Chi-square (CH)	13926	75.63%	92.87%	0.8337	85.78%	59.21%	0.7006	78.59%	0.7671
F-score	6380	75.61%	91.99%	0.8300	84.80%	59.73%	0.7009	78.30%	0.7655
Odds Ratio (OR)	6963	75.47%	91.21%	0.8260	83.46%	59.61%	0.6955	77.80%	0.7607
Mutual Information (MI)	55904	72.12%	93.56%	0.8145	85.53%	50.72%	0.6368	75.38%	0.7257

Table 7: Liberal vs. Conservative classification results by different feature selection methods (L_P, L_R, L_F1, R_P, R_R, and R_F1 denote precision, recall and F1 of liberal and conservative blog prediction respectively)

improvement).

4.1.3 Discriminating words To examine if feature selection helps to identify ‘discriminating words’ used by either group of bloggers, we listed the top 10 words by liberal bloggers and the top 10 words by conservative bloggers found by each feature selection method in Table 16.

From the table, however, it can be seen that many of the top ranked words are in fact partial names of top ranked bloggers (e.g., *fredoglake*). Further, when we looked down the list beyond the top 10 words, many ‘words’ turned out to be non-valid English words (e.g. *superwealthy*). Although these ‘words’ helped classifiers, they do not provide insight into what topics bloggers with different political leanings speak of or what opinions they tend to hold. In order to find out these insights, we removed ‘words’ that were partial bloggers’ names and that did not appear in WordNet[12]. Unfortunately, we found again that they did not provide clear clues. As a result, we manually examined all top 7000 words³ in both liberal and conservative blogs. It showed the following observation: meaningful words that distinguished both topics and opinions were not ranked consistently high and were intervened by non-politically-related words. In our opinion, the reason for this ranking of top words is two-fold: first, bloggers often exhibit free writing styles such as using non-standard acronyms and tolerating typos; second, blog postings tend to be on multiple topics. As a result, some invalid English ‘words’ appeared very often and got ranked higher. In addition, words that were irrelevant to politics were frequent and were thus ranked towards the top as well.

Though the list is not as what we expected, we did find some interesting observations. Firstly, we looked for words by seven categories that are often used to evaluate opposite political leanings. By examining the results obtained, we found that they showed some differences

between liberal and conservative bloggers. For example, “healthcare” related words, such as *healthcare*, *drug* and *medicate* only appeared in the top 7000 in liberal blogs. Similarly, words related to “government” like *legislation* and *regulatory* also appeared among the top 7000 in solely liberal blogs. In contrast, *immigration*, *border*, *abortion* and *homosexual* only appeared in the top 7000 in conservative blogs. Table 8 lists some of these words.

Category	Words	Liberal	Conservative
Social Security	retirement	×	✓
	welfare	×	✓
Healthcare	healthcare	✓	×
	drug	✓	×
	medicare	✓	×
Government	legislation	✓	×
	regulatory	✓	×
Immigration	immigration	×	✓
	border	×	✓
Abortion	abortion	×	✓
Gay&Lesbian	homosexual	×	✓
	heterosexual	✓	×
Economy	inflation	×	✓
	unemployment	×	✓
	deregulation	✓	×
	protectionism	×	✓
	monopolize	✓	✓
	welfare	×	✓
Religion	religion	×	✓
	charity	×	✓
	astrology	✓	×
Other	democratic	✓	×
	republican	✓	×
	liberal	✓	✓
	conservative	✓	✓

Table 8: Notable words from the top 7000 ‘discriminating words’ by categories: Liberal vs. Conservative

Secondly, we looked for meaningful words that either appeared only among the top 7000 liberal blogs or appeared only among the top 7000 conservative blogs and that we thought might be relevant to different political leanings. Some of these words are listed in Table 9.

³The reason that we chose top 7000 words in either category is due to the results that the best average F1 was achieved at this size of features.

Liberal
overpriced militarism plutocracy monopolize democratization patriarchy neoliberalism modernist sectarianism corporatism cronyism consumerism anarchism colonialism recidivism fatalism theism reformism materialism symbolism activism unilateralism sadism centrism cynicism feminism deism minimalism ageism futurism communalism egoism televangelism zoroastrianism sexism republicanism mormonism pragmatism idealism leninism feudalism imperialism
Conservative
surrealism collectivist protectionism multiculturalism leftism romanticism stalinism collectivism escapism rationalism commercialism postmodernism paternalism rightism moralism wahhabism ostracism

Table 9: Notable words from the top 7000 ‘discriminating words’: Liberal vs. Conservative

4.1.4 Using out-link To see if out-link improves performance, we included all eight out-link features with optimal numbers of BoW features found by five feature selection methods. Results indicated a consistent improvement in all cases, as shown in Table 10 (aggregation method is ‘sum’ for non-F-score feature selection). On average, a 5% and a 6.5% improvement were seen on overall accuracy and average F1 respectively.

4.1.5 Handling imbalanced data sets From Table 6 and Table 7, low recall of conservative blogs (less than or around 60%) and low precision of liberal blogs (around 75%) showed the prediction of political leanings was biased in that more conservative blogs were mistakenly classified as “liberal”. This can also be seen in Table 10, though recall of “conservative” and precision of “liberal” improved a lot. Our first guess was that this might be caused by the fact that more liberal blogs were used in training. As a result, our data set is imbalanced: more positive class instances (‘liberal’) than negative class instances (‘conservative’). To verify if this was the case, we created three different balanced data sets: 1) “top793”, which consisted of all conservative blogs and 793 top liberal blogs; 2) “bot793”, which consisted of all conservative blogs and 793 bottom liberal blogs; and 3) “random793”, which consisted of all conservative blogs and 793 randomly selected liberal blogs. The classification results of using all BoW features are shown in Table 11.

From the results of “random793”, we saw balanced performance as a result of a balanced data set. But

from the first two balanced data sets, “top793” and “bot793”, we still saw somewhat imbalanced results: more liberal blogs from top 793 were classified as ‘conservative’ while more conservative blogs were classified as ‘liberal’ on “bot793” data set. So, this indicates that an imbalanced data set caused by different number of ‘positive’ and ‘negative’ instances is not the only reason that leads to an imbalanced performance. The differences between what constructs positive class instances and what constructs negative class instances should be other important factors. We termed this type of differences “imbalanced blog content”. However, it remains unknown what these differences really are, which should be the goal of our next work.

Learning from imbalanced data sets has recently become a hot research area, in which the data sets under study share the same property: the distribution of classes is highly unbalanced, some classes have far more instances than others[20, 4]. Previous work on this problem can be roughly grouped into three categories: the first is to skew the training data distribution to favor a small class, the second is the design of new algorithms or modifications to traditional algorithms to address this problem, and the third is to do a feature selection to favor small class classifications.

In our work, the distribution of positive class and negative class is not highly unbalanced. In addition, the size of the whole corpus is not very big. Because of this, we wanted to use all the instances available to us without skewing the training data. On the other hand, as of this point, it remains unknown what “imbalanced blog content” really is. As a result, it is difficult for us to use feature selection to favor small class classification. With these considerations, to remedy both types of ‘imbalance’ (uneven distribution of liberal and conservative blogs in training set and differences in content of either category of blogs) in order to achieve a balanced classification performance, our solution is to balance the cost of ‘false positive’⁴ error and the cost of ‘false negative’⁵ error. This method is known as ‘cost-sensitive learning’[7, 11, 10, 40], and recent work has shown its success in dealing with imbalanced class problems[29, 38, 27]. In *SVM^{light}*, this can be achieved by tuning ‘cost factor’[32] to the following value: the number of conservative blogs over the number of liberal blogs. In our case, we have 1054 liberal blogs and 793 conservative ones, as a result, the value of ‘cost factor’ is 0.75. This way, we give more weight to adjusting errors on negative instances (‘conservative’ blogs), which helps classify more negative instances correctly. As a result,

⁴Conservative blogs are classified as ‘liberal’

⁵Liberal blogs are classified as ‘conservative’

Feature selection	BoW size	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	\bar{F}_1
Odds Ratio	6963	80.70%	93.26%	0.8653	88.60%	69.66%	0.7800	83.25%	0.8226
F-score	6380	80.35%	93.07%	0.8624	88.15%	68.99%	0.7741	82.85%	0.8182
Chi-square	13926	79.66%	93.56%	0.8605	88.63%	67.29%	0.7620	82.41%	0.8128
Information Gain	6963	79.59%	93.16%	0.8584	87.96%	67.56%	0.7642	82.29%	0.8113
Mutual Information	55904	75.32%	94.43%	0.8380	88.87%	57.88%	0.7010	78.92%	0.7695

Table 10: Liberal vs. Conservative classification (BoW + out-link) (L_P, L_R, L_F1, R_P, R_R, and R_F1 denote precision, recall and F1 of liberal and conservative blog prediction respectively)

Data set	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	\bar{F}_1
top793	88.23%	72.75%	0.7974	77.05%	90.17%	0.8310	81.46%	0.8142
bot793	72.80%	92.43%	0.8145	89.79%	65.08%	0.7546	78.76%	0.7846
random793	80.75%	81.86%	0.8130	81.97%	80.21%	0.8108	81.03%	0.8119

Table 11: Balanced data sets: Liberal vs. Conservative classification (all BoW features) (L_P, L_R, L_F1, R_P, R_R, and R_F1 denote precision, recall and F1 of liberal and conservative blog prediction respectively)

more ‘conservative’ blogs can be correctly classified. In Table 12, results of using different ‘cost factor’ values are listed by using all BoW features. It can be seen that the highest average $F1$ is achieved when ‘cost factor’ is 0.75. The row corresponding to the ‘cost factor’ of 1 indicates the experiment in which no remedy was actually used in handling imbalanced data sets problems. So, it shows the same results as Table 6. When ‘cost factor’ is assigned a higher value, more ‘liberal’ blogs are detected at the cost of mis-classifying more ‘conservative’ ones; on the other hand, with lower ‘cost factor’ values, more blogs are classified as ‘conservative’ due to the mis-classification of ‘liberal’ blogs instead. Clearly, the performance at the ‘cost factor’ of 0.75 achieved a significant improvement over no cost adjustment.

To test if the improvement of balancing errors of both classes is consistent, we set ‘cost factor’ to 0.75 in the political leaning classification by using a combination of all BoW and out-link features. As can be seen in Table 13, results show significant improvement over using all BoW features with the same ‘cost factor’ value.

Category	Precision	Recall	F1
Liberal	89.51%	86.81%	0.8814
Conservative	83.36%	86.39%	0.8485
Average F1:		0.8650	
Overall accuracy:		86.63%	

Table 13: Using ‘Cost-factor’ of 0.75: Liberal vs. Conservative classification (all BoW features + out-link)

4.2 Political blog classification Firstly, we used all BoW features to train SVM^{light} classification model

without feature selection. 1054 liberal blogs and 793 conservative blogs were combined to generate positive class instances, that is, political blog sets. 2809 non-political blogs were used to generate negative class instances. Then all instances of both classes were combined to do a 10-fold cross-validation. Table 14 shows results.

Category	Precision	Recall	F1
Politics	92.98%	85.65%	0.8917
Non-Politics	90.46%	95.44%	0.9288
Average F1:		0.9102	
Overall accuracy:		91.40%	

Table 14: Political vs. Non-Political (all BoW features)

Next, using the top 100 liberal and the top 100 conservative blogs as positive classes and 200 non-political blogs (selected evenly from 41 non-political categories) as negative classes, by applying feature selection, similar results were seen as in political leaning categorization. Due to space constraints, we will not show them. Instead, we will only show the top 20 ‘politically-charged words’ ranked by F-score method in Table 15 (similar words were not discarded).

bush	political	democratic	politics	democrats	iraq
war	campaign	president	republican	election	liberal
republicans	congress	clinton	government	military	
hillary	presidential	america			

Table 15: Top 20 ‘politically-charged words’ (by F-score)

Finally, we tested performance by combining eight out-link features with selected BoW features. Unlike political leaning categorization, however, inclusion of

Cost-factor	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	\bar{F}_1
1.4	72.07%	97.53%	0.8289	93.90%	49.45%	0.6479	76.89%	0.7384
1.2	73.19%	97.25%	0.8352	93.66%	52.35%	0.6716	77.97%	0.7534
1	73.82%	94.13%	0.8250	87.71%	54.94%	0.6757	77.20%	0.7503
0.8	82.09%	88.24%	0.8505	82.85%	74.03%	0.7819	82.14%	0.8162
0.75	86.01%	81.87%	0.8389	77.50%	81.98%	0.7968	81.92%	0.8179
0.6	93.02%	62.72%	0.7492	65.56%	93.70%	0.7715	76.62%	0.7603

Table 12: Using Cost-factor: Liberal vs. Conservative classification (all BoW features) (L_P, L_R, L_F1, R_P, R_R, and R_F1 denote precision, recall and F1 of liberal and conservative blog prediction respectively)

out-link features did not achieve a consistent increase in performance. In several cases, it even caused slight decrease, though both increase and decrease were not significant.

4.3 Blog filtering As we mentioned in section 3.1, we manually discarded some “noise” blogs and used the remaining blogs for training and testing. To ensure that this does not bring over optimistic results, in this subsection, we claim that this removal has little impact on the experimental results reported above.

Firstly, blogs with small number of texts on their *front page* will not be classified at all. We rely on the text on the *front page* of a blog to classify it into different categories. In this regard, blogs that specialize in mostly providing image, music and video on their web sites are out of our consideration. Other than these blogs, if there are few words on the *front page* of a blog, it indicates that its owner does not conduct updates often. We thus feel it is reasonable to deem these blogs unimportant and we do not take them into consideration for categorization.

Secondly, non-English blogs can be automatically detected with good precision. Though it is not trivial to perform natural language identification, it is more accurate to detect if a body of text is in English or not. To test this, we performed two experiments by using an open-source Perl module `Lingua::Identify`⁶. One is to test if a “noise” blog in non-English is included as candidate for classification. For this purpose, we gathered all 132 non-English blogs previously identified and discarded manually. Then, we parsed and extracted only the text from their *front pages* and passed them to `Lingua::Identify` to identify their languages. Only 10 of these blogs were mis-identified as English. Then, we passed these 10 blogs to political blog classifier and only two of them were classified as political. In the other experiment, we tested if blogs written in English will be mistakenly identified as non-English if we first

perform a language identification. For this purpose, we identified the languages of all 1054 liberal blogs and 793 conservative blogs by using `Lingua::Identify`. Only 28 blogs out of them were deemed as non-English, with a low error rate of 1.41%. We think this is acceptable since only two out of every 100 blogs are excluded from categorization.

Lastly, it is our claim that a portal web site that only displays a listing of categories should not be taken as a candidate of either liberal or conservative category. To see if our political classification model is able to exclude such portal blogs, we fed into the model 102 portal blogs that were discarded when we were building the corpus and it only classified 15 as political. Then we read all these 15 blogs and found that most of them contain some ‘politically-charged words’. For example, in such a blog⁷, a lot of words like *Afghanistan War*, *Anti War*, *gun control*, and names of all presidential candidates, etc. appeared in the listings. This shows that our classifiers have good discerning capability to block certain blog spams, which are usually web hosting sites with listings of advertisement on their front pages.

4.4 Comparison to similar work Most related work similar to ours are [33, 28] and [9]. Both aimed to categorize blogs in terms of two opposite political leanings as well. In [33, 28] though, blogs considered were politics discussion postings, while in [9], blogs were restricted to topics related to ‘President Bush’ and ‘Iraq War’. [33, 28] applied sentiment-analysis-based method, Naive Bayesian classification method (on unigram) and text clustering method with overall accuracy achieved between 41% ~ 73%. [9] combined unigram with polarity (i.e. if a word is after a negative word like *not*, the former is marked as a negative feature). They applied both Navie Bayes and SVM classifiers, which achieved an average accuracy of 78.06% and 75.47% respectively. They tested classification with a varied number of features. However, it was not reported if this was done with some feature selection methods.

⁶<http://search.cpan.org/author/COG/Lingua-Identify-0.19/lib/Lingua/Identify.pm>

⁷<http://wvkossacks.blogspot.com>

In our work, we did not compare SVM to other classification algorithms. Instead, a baseline can be a trivial algorithm simply based on a blind guess, which can obtain an accuracy of 57% in political leaning classification and 60% in political vs. non-political blog classification. We explicitly applied feature selection. In addition, we used links in classification and saw a consistent improvement in political leaning classification. We also investigated the problem of classifying a blog as political vs. non-political. Another important task that we performed is partly tackling the imbalanced dataset problem. By treating false positive error and false negative error differently, we were able to achieve balanced performance on classifying both categories of political leanings and, as a result, achieve higher overall accuracy. Lastly, we rely on the *front page* of a blog for classification. Generally, however, we do not restrict blogs except in ignoring certain “noise” blogs.

4.5 Discussion Firstly, from political vs. non-political blog classification results, it shows that ‘bag of words’ alone can achieve a fairly good performance. In addition, by using feature selection, important ‘politically-charged words’ can be found (see Table 15).

In liberal vs. conservative classification, using only ‘bag of words’ on our imbalanced corpus can only obtain an overall accuracy below 80% (see Table 6 and Table 7) and performance is biased. By adjusting the costs of false positive errors and false negative errors, not only is the performance of both political leaning categories balanced, the overall accuracy is improved significantly as well, with 81.92% at ‘cost factor’ of 0.75.

Effectiveness of including out-link features (see Table 10 and Table 13) in political leaning categorization reveals there is a somewhat clear pattern in citation behavior among political bloggers. It was also reported in [1] that both liberal and conservative bloggers tended to cite each other within the community while cross citing accounted for less portion of total links. In our corpus, inter-category citation links only accounted for 8% of both intra- and inter-category links (links to external sites were not considered).

Lastly, a slight difficulty in finding meaningful ‘discriminating words’ in political leaning categorization shows the weakness of using solely ‘bag of words’ features due to the free writing style of bloggers and the multi-topic property of blog postings. In addition, by examining top ranked words, it also indicates that the ‘bag of words’ plays a smaller role in detecting differences of opinion between liberal and conservative bloggers since few such opinion-words were seen among them.

5 Conclusions

In this paper, we have discussed political blog categorization problems and investigated the effectiveness of using ‘bag of words’ and some link features in training SVM classifiers. Preliminary experiments showed a good performance in political vs. non-political blog classification. In political leaning classification, the best overall accuracy achieved was around 86% with the help of cost-sensitive learning methods. Finally, feature selection revealed some limitations of using ‘bag of words’ features in political leaning classification.

In future work, we would like to improve political blog categorization based on the following considerations. Firstly, we want to manually analyze blog postings and design appropriate preprocessing techniques to clean raw texts. Secondly, we will use front pages spanning different lengths of period to analyze temporal issues that may influence the performance of models thus built. Thirdly, we plan to apply models to new blogs and analyze results to evaluate combinations of both techniques discussed here. Next, we will analyze “imbalanced blog content” to study the factors that cause imbalanced political leaning categorization performance against a balanced training set that contains an equal number of liberal and conservative blogs. Lastly, we want to explore new features that will be more effective than current ones and will be able to reveal words that are characteristic of a specific political orientation.

References

- [1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery, LinkKDD '05*, pages 36–43, 2005.
- [2] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12(9), September 2007.
- [3] K. Balog and M. de Rijke. Decomposing bloggers’ moods: Towards a time series analysis of moods in the blogosphere. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, at WWW2006*, 2006.
- [4] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1):1–6, 2004.
- [5] Y.-W. Chen and C.-J. Lin. *Combining SVMs with various feature selection strategies*, chapter 12, Feature extraction foundations and applications, pages 315 – 325. Springer, 2006.
- [6] H.-J. Choi and M. S. Krishnamoorthy. Categorization of blogs through similarity analysis. *Intelligence and Security Informatics, 2007 IEEE*, pages 160–165, 2007.

Feature selection	Liberal	Conservative	Aggregated('sum')	Aggregated('max')
Information Gain (IG)	footballs firedoglake wizbang pandagon crooks hugh liars digby hewitt spades	firedoglake pandagon crooks eschaton skippy marsh majikthise liars oasis tbogg	firedoglake pandagon crooks liars eschaton skippy majikthise marsh footballs tapped	firedoglake pandagon crooks eschaton skippy marsh majikthise liars footballs oasis
Chi-square (CH)	firedoglake pandagon crooks liars skippy eschaton majikthise tapped progressive marsh	firedoglake pandagon crooks liars skippy eschaton majikthise tapped progressive marsh	firedoglake pandagon crooks liars skippy eschaton majikthise tapped progressive marsh	firedoglake pandagon crooks liars skippy eschaton majikthise tapped progressive marsh
F-score			firedoglake pandagon crooks digby liars skippy eschaton majikthise tapped mahablog	
Odds Ratio (OR)	archduke chopped shutup bikinis medi- anews 7344184 butt- load 27702 magnifies randomordercontent	s2248 squelch bots wieczorek coolwiz fearfully holies reagan21 psyche profanity	firedoglake pandagon marsh footballs oasis tbogg coaster sideshow echidne wizbang	firedoglake pandagon marsh footballs oasis tbogg coaster sideshow echidne wizbang
Mutual Information (MI)	archduke chopped shutup bikinis medi- anews 7344184 butt- load 27702 magnifies randomordercontent	s2248 squelch bots wieczorek coolwiz fearfully holies reagan21 psyche profanity	s2248 squelch arch- duke chopped bots shutup medianews bikinis coolwiz wieczorek	s2248 squelch archduke chopped bots shutup medianews bikinis coolwiz wieczorek

Table 16: Top 10 discriminative words by Liberal and Conservative bloggers

- [7] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.
- [8] Y. Dourisboure, F. Geraci, and M. Pellegrini. Extraction and classification of dense communities in the web. In *Proceedings of WWW'07*, pages 461–470, 2007.
- [9] K. T. Durant and M. D. Smith. Mining sentiment classification from political web logs. In *Proceedings of WEBKDD'06*, 2006.
- [10] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
- [11] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. Adacost: misclassification cost-sensitive boosting. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 99–105, 1999.
- [12] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [13] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of KDD*, pages 150–160, 2000.
- [14] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proceedings of HYPERTEXT '98*, pages 225–234, 1998.
- [15] K. E. Gill. How can we measure the influence of the blogosphere? In *Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*, 2004.
- [16] J. Graf. The audience for political blogs: New research on blog readership, 2006. Available as [http://www.ipdi.org/UploadedFiles/Audience for Political Blogs.pdf](http://www.ipdi.org/UploadedFiles/Audience%20for%20Political%20Blogs.pdf).
- [17] S. Han, Y. yeol Ahn, S. Moon, and H. Jeong. Collaborative blog spam filtering using adaptive percolation search. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, at WWW2006*, 2006.
- [18] H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *Proceedings of WWW*, pages 661–669, 2005.
- [19] K. Ishida. Extracting latent weblog communities: A partitioning algorithm for bipartite graphs. In *Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem, at WWW2005*, 2005.
- [20] N. Japkowicz. AAI tech report ws-00-05. In N. Japkowicz, editor, *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*, 2000.
- [21] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.
- [22] H. KOICHI, F. NORIYA, and T. JUNICHI. Text categorization for japanese blog entries. Technical report, IPSJ SIG Technical Reports, 2005.

- [23] P. Kolari, T. Finin, and A. Joshi. Svms for the blogosphere: Blog identification and splog detection. In *Proceedings of AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
- [24] M. Koppel, J. Schler, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. In *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
- [25] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, May 2003.
- [26] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of blog communities based on mutual awareness. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, at WWW2006*, 2006.
- [27] X.-Y. Liu and Z.-H. Zhou. The influence of class imbalance on cost-sensitive learning: An empirical study. In *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*, pages 970–974, 2006.
- [28] R. Malouf and T. Mullen. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW)*, 2007.
- [29] D. Margineantu. When does imbalanced data require more than cost-sensitive learning? In *Proceedings of Workshop on Learning from Imbalanced Data, National Conference on Artificial Intelligence (AAAI-2000)*, pages 47–50, 2000.
- [30] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of AIRWeb '05, at WWW 2005*, 2005.
- [31] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.
- [32] K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 268–277, 1999.
- [33] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 159 – 162, 2006.
- [34] K. Narisawa, Y. Yamada, D. Ikeda, and M. Takeda. Detecting blog spams using the vocabulary size of all substrings in their copies. In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem, at WWW2006*, 2006.
- [35] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
- [36] S. Shulman, J. Callan, E. Hovy, and S. Zavestoski. Language processing technologies for electronic rule-making: A project highlight. In *Digital Government Research*, pages 87–88, 2005.
- [37] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP*, 2006.
- [38] G. M. Weiss. Mining with rarity - problems and solutions: A unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.
- [39] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, 1997.
- [40] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 435–442, 2003.
- [41] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *Proceedings of WWW*, pages 221–230, 2007.