Political Leaning Categorization by Exploring Subjectivities in Political Blogs

Maojin Jiang and Shlomo Argamon
Linguistic Cognition Lab, Department of Computer Science, Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616 USA

Abstract—*This paper addresses a relatively new text categorization problem: classifying a political blog as either 'liberal' or 'conservative', based on its political leaning. Instead of simply using "Bag of Words" features (BoW) as in previous work, we have explored subjectivity manifested in blogs and used subjectivity information thus found to help build political leaning classifiers. Specifically, our subjectivity based approach is two fold: 1) we identify subjective sentences that contain at least two strong subjective clues based on the General Inquirer dictionary; 2) from subjective sentences identified, we extract opinion expressions and BoW features to build political leaning classifiers. Experiments with a political blog corpus we built show that by using features from subjective sentences can significantly improve the classification performance. In addition, by extracting opinion expressions from subjective sentences, we are able to reveal opinions that are characteristic of a specific political orientation to some extent.*

Keywords: political leaning categorization, subjectivity analysis, opinion expression, text mining

1 Introduction

As more political blogs go online, it is both interesting and useful to locate the bloggers' positions along the political leaning continuum, from liberal to conservative.

Though text categorization has been heavily researched, previous work on political blog categorization in terms of political leanings has been limited. It is known that political blogs are highly opinionated and rich in subjective languages. Intuitively, it is the bloggers' different opinions on common issues that mark a border between liberal blogs and conservative ones. This makes the political leaning categorization both different and more difficult than traditional text categorization applications that have been researched. Therefore, it has motivated us to explore subjectivity manifested in political blogs and use subjectivity information thus found to help build political leaning classifiers.

In the following, we will first give an overview of previous work in related areas. Then, we will introduce our corpus, basic framework of our approach and evaluation methods. Next, we will describe our approaches to detecting subjectivity and using what we find to build political orientation classifiers in detail, with experimental results showing improvement over previous work. Lastly, we will discuss our methods and give future work plans.

2 Related work

2.1 Political blog analysis

Some work has been seen in analyzing political blogs in recent years. For example, 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. Some 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 [2], authors tried to classify political orientations in online political discussion postings. Similarly, [3] performed sentiment classification on blog postings relevant to certain topics.

2.2 Blog categorization

Blog (or Blogger) categorization has become a 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 [4], [5], [6], [7], [8]. [9], [10] explored different methods in text categorization of blogs in terms of similarity of topics in general. [11], [12] analyzed significant differences in writing styles and contents 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. [13], [14] identified bloggers' moods by capturing salient linguistic features with the help of machine learning algorithms. In [2], [3], authors aimed to classify certain blog postings into categories of opposite political orientations.

Lastly, as a result of emerging blog spam, blog spam detection and filtering has also become popular recently [15].

2.3 Subjectivity analysis

Subjectivity is defined as "judgment based on individual personal impressions and feelings and opinions rather than external facts" in WordNet [16]. From this definition, it is clear that 'subjectivity' is distinguished from factual information. According to [17], 'subjectivity' in natural language refers to aspects of language used to express opinions, evaluations, and speculations.

Recently, due to the popularity of sentiment analysis of online product reviews (for example, movie reviews as in

[18]), it becomes an important task to separate subjective language from text that only describes actual facts. As a result, subjectivity analysis plays an important role in such applications as review analysis [19], [20], information extraction [21], flame recognition [22] and information retrieval [23], [24], among others.

Methods used to build a subjectivity vs. objectivity classifier or to identify subjective language in texts can be roughly classified into three groups: rule-based, theory-based and learning-based. Rule-based methods rely on a list or a dictionary of subjective clues to determine subjective text [25]. Theory-based methods [19], [26] are mainly based on Appraisal Theory developed by [27], based on the foundation of Systemic Functional Linguistics [28]. Lastly, learning-based methods apply learning algorithms on both labeled and unlabeled data to learn classifiers, patterns or subjective clues [29], [17], [25].

3 Political leaning categorization as a classification task

Our approach to political blog categorization in terms of political leanings is to use state-of-the-art machine learning algorithms to learn a classifier that can determine political leanings of political blogs and classify them as either *liberal* or *conservative*.

To learn a classifier, we first created a political blog dataset that contains *front pages* of blogs explicitly labeled with political orientations. 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".

Then the corpus is divided into a training set and a testing set. Each blog is represented as a vector of values of the features selected. In addition, *liberal* blogs are labeled as 'positive' classes while *conservative* blogs as 'negative' classes. Next, building classifiers are conducted by using standard machine learning techniques that have been successfully used for text categorization tasks in the past on the training set. Last, performance of the classifiers are evaluated on the testing set.

In the remaining part of this section, we will first introduce our corpus, followed by machine learning algorithm we used. Then, we will briefly mention the non-subjectivity-related features that we have used in previous work. Next, we will describe the metrics that are commonly used in evaluation of text categorization methods. Lastly, previous classification results will be reported here again so that we can evaluate the effectiveness of our new methods by comparison.

3.1 Corpus

Our political blog corpus consists of *front pages* of two groups of blogs, namely, 1054 liberal blogs and 793 conservative blogs, totalling more than 200MB. The URLs of these blogs were obtained from five blog catalog websites[1]. 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". All these blogs are written in English and each of them has no less than 140 words on its *front page* (excluding HTML tags). In addition, the corpus contains no duplicate blog and no blog appears as both liberal and conservative.

3.2 Classifier learning

In the past, Support Vector Machine learning algorithm (SVM) has been successfully used in text categorization. In this paper, we report on results using classification in Joachims' SVM^{light} implementation of SVM [30]. This tool can handle problems with thousands of support vectors efficiently, has scalable memory requirements, and has many useful options for learning (for example, allowing to conduct cost-sensitive learning by adjusting the value of 'cost-factor' option). Unless otherwise stated, all results reported in this paper were obtained by using SVM^{light} with linear kernel in a 10-fold cross-validation by setting 'cost-factor' to 0.75 (calculated as the number of conservative blogs divided by the number of liberal blogs, i.e., $\frac{793}{1054}$) to balance the cost of 'false positive'[2] error and the cost of 'false negative'[3] error, as a solution to imbalanced dataset problem (in our corpus, more liberal blogs than conservative ones used in learning) [31].

3.3 Non-subjectivity features

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.

3.4 Evaluation methods

Traditional evaluation metrics in text categorization are used in evaluating political leaning 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* 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

[1] blogcatalog.com, blogarama.com, etalkinghead.com, campaignsandelections.com, and blogs.botw.org

[2] Conservative blogs are classified as 'liberal'.

[3] Liberal blogs are classified as 'conservative'.

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. In the following, 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.

3.5 Preliminary results by using non-subjectivity features

In our previous work on the problem [31], the highest overall accuracy of political leaning classification by only using BoW features from the full text in every blog was 81.92%.

4 Subjectivity-based approach

Our idea to explore subjectivity in political blogs is two fold: first, we use a rule-based method to detect subjective sentences in blogs; second, we exploit subjectivity features from subjective sentences identified to build political leaning classifiers and conduct classification. In the following, we will describe our approach in detail.

4.1 Identifying subjective sentences

We use a rule-based subjective classifier to separate subjective sentences from objective ones. The rule of the subjective classifier is simple: if a sentence contains at least two strong subjective clues, it is classified as subjective.

To look for strong subjective clues, we chose the General Inquirer dictionary[4] (GI) due to its abundant semantic categories. In total, GI has 182 categories in 26 groups, mainly of nouns, verbs, adjectives and adverbs. For our purpose, we chose the following nine categories in GI. They are "Strong", "Hostile" and all seven categories in category group No. 3 labeled as *Words of pleasure, pain, virtue and vice*, including, "Pleasur", "Pain", "Feel", "Arousal", "EMOT", "Virtue" and "Vice" (for descriptions and words of these categories, please visit the General Inquirer website[5]). Thus, all words belonging to any of these nine categories are considered as strong subjective clues. The Table I lists all nine categories with the number of words belonging to them. In total, there are 4004 different entries (an entry is a word sense) in GI that appear on our strong subjective clue list.

TABLE I
GI CATEGORIES OF STRONG SUBJECTIVE CLUES

Category	Size	Category	Size	Category	Size
Strong	1902	Hostile	833	Pleasur	168
Pain	254	Feel	49	Arousal	166
EMOT	311	Virtue	719	Vice	685

So, to determine subjectivity of a sentence, we need to look up a word in the sentence in GI. In GI, a word usually belongs to multiple categories. For some words, GI contains more than one sense of them, each with a list of categories to which the sense belongs. In addition, most categories consist of words that have multiple parts-of-speech. As a result, to accurately obtain a list of categories for a word in a sentence, a full disambiguation method is required. In our current work, we adopt a simple disambiguation method that bases on the part-of-speech (POS) of a word. Specifically, for each entry in GI, which is a word sense, we determine its POS from the following categories: 'Othtags', 'Defined' and several explicit category labels for verbs and adjectives. Then, given a POS-tagged word, we assign it to all categories of the same word with the same POS in GI if found. For words not in GI, we do not assign categories.

To sum up, the whole process to classify a sentence in a blog as *subjective* or *objective* works as follows. First, for a HTML file of a blog, HTML tags and some irrelevant tagged text (e.g. text tagged as *script* and *comment*) are discarded. Then, the remaining text is split into sentences. Next, each sentence is parsed by a POS tagger and each word is reduced to its base form by a morphological analyzer of English, after which the sentence is fed into the subjective classifier. For each POS-tagged word in its base form, the classifier first looks up its category in GI. If a word belongs to either of the nine categories as mentioned above, it is determined as a *strong subjective clue*. If two such clue words are found, the classifier stops examining more words and classifies the sentence as *subjective*. If a sentence is not classified as *subjective*, it is labeled as *objective*.

By applying the subjective classifier to both liberal and conservative blogs in our corpus, around 63% sentences in both liberal and conservative blogs were classified as *subjective*. Since we have not manually labeled any sentence as subjective or objective in our corpus, we do not directly evaluate this result. Instead, we rely on performance of using these subjective sentences in political leaning categorization as extrinsic evaluation, which we will discuss in the next subsection.

4.2 Using subjective sentences in political leaning categorization

To evaluate the effectiveness of exploring subjectivity in political leaning categorization, we performed experiments by using different features solely from subjective sentences, as described in the following.

1) BoW: First, we compared political leaning categorization results by using BoW from all texts to using BoW from only subjective sentences. In addition, to compare GI to other strong subjective clues that have been successfully used in subjectivity analysis, we selected such a 'Subjectivity Lexicon'[6] generated as described in [32]. In the following discussion, we will use 'WW' to refer to it. In WW, 5569 entries (a word-POS pair with other labels, e.g. polarity) are labeled as 'strongsubj'. So, in a subjective classifier by using WW as strong subjective clue list, only a word that matches

[4] http://www.wjh.harvard.edu/∼inquirer/

[5] http://www.wjh.harvard.edu/∼inquirer/homecat.htm

[6] http://www.cs.pitt.edu/mpqa

TABLE II

Political leaning classification (BoW): with vs. without identifying subjective sentences

Classifier	Size (%)	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	Avg. F1
all_{BoW}	100	86.01%	81.87%	0.8389	77.50%	81.98%	0.7968	81.92%	0.8179
$GIsubj_{BoW}$	63	87.65%	82.63%	0.8507	78.86%	84.13%	0.8141	**83.28%**	**0.8324**
$GIobj_{BoW}$	37	75.54%	70.01%	0.7267	63.79%	69.85%	0.6669	69.94%	0.6968
$WWsubj_{BoW}$	25	86.82%	75.81%	0.8094	72.73%	84.50%	0.7817	79.54%	0.7956
$WWobj_{BoW}$	75	83.59%	71.81%	0.7725	68.68%	81.09%	0.7438	75.80%	0.7581
$rand25_{BoW}$	25	78.33%	67.18%	0.7233	63.51%	75.16%	0.6884	70.60%	0.7059
$rand63_{BoW}$	63	81.66%	74.95%	0.7816	70.37%	77.31%	0.7368	75.96%	0.7592

the word-POS pair labeled as ‘strongsubj’ in WW is deemed as a strong subjective clue. By applying this new subjective classifier to our corpus, only around 25% sentences in both liberal and conservative blogs were classified as *subjective*.

The Table II shows results by using different political leaning classifiers, where ‘$GIsubj_{BoW}$’ and ‘$GIobj_{BoW}$’ are the ones using BoW features from subjective sentences and objective sentences detected by using GI dictionary respectively; similarly, ‘$WWsubj_{BoW}$’ and ‘$WWobj_{BoW}$’ are the ones using WW list; and ‘all_{BoW}’ means the one that all sentences are used; the column ‘Size’ indicates percentage of all sentences used in experiments. It can be seen that ‘$GIsubj_{BoW}$’ achieves the best performance and it outperforms ‘all_{BoW}’ on overall accuracy by 1.66% ($p < 0.005$). Unfortunately, ‘$WWsubj_{BoW}$’ performs worse than ‘all_{BoW}’ with a decrease of 2.90% ($p < 0.05$) on overall accuracy. This indicates that GI is more effective than WW in our political leaning analysis. Therefore, we will only use GI-based subjective classifier in the following subjectivity analysis and use the subjective sentences classified by it in building political leaning classifiers. From the Table II, it also indicates that classifiers by using objective sentences all perform much worse than their subjective counterparts and ‘all_{BoW}’. In addition, to compare ‘$GIsubj_{BoW}$’ and ‘$WWsubj_{BoW}$’ to classifiers that are trained on the same number of sentences, we randomly selected 25% and 63% sentences and built classifiers from them. The rows in the table, ‘$rand25_{BoW}$’, and ‘$rand63_{BoW}$’ correspond to these cases respectively. Clearly, none of them can beat either ‘$GIsubj_{BoW}$’ or ‘$WWsubj_{BoW}$’.

2) Opinion Expressions: A direct benefit of detecting subjective sentences is to make opinion detection more easier since these sentences are often rich in sentiment or opinion expressions. So, our next goal is to extract useful opinion expressions and use them as binary features in building political leaning classifiers that are expected to have a higher classification accuracy.

In our previous work on semantic analysis of political blogs [33], we observed that semantics in political blogs in our corpus can be mostly explained by verbs and nouns. It also suggested that features beyond single words (such as collocations of verbs and nouns, together with adjectives and adverbs as modifiers) be used to create political leaning classifiers. However, instead of blindly allowing an arbitrary combination of words of these four parts-of-speech as a valid opinion expression, we should take more factors into considerations. First, in our opinion, an opinion expression should contain at least a noun, which represents an object that an opinion is about. Then, in addition to nouns, an opinion can be expressed by using either verbs or adjectives or their combinations. If there is a verb in an opinion expression, it might have one or more adverbs as modifiers. In addition, as people often either support or oppose to something like policies, it would be desirable to know if a blogger holds *positive* or *negative* attitude in an opinion expression. In GI, we have categories ‘Positiv’ and ‘Negativ’ that correspond to these two stances or two values of *Orientation* in appraisal theory, respectively[7]. Lastly, considering that people tend to use negation to strengthen their attitudes in natural language, we also include the GI category ‘Negate’[8] in opinion expression extraction, which is termed as ‘marked’ *Polarity* in appraisal theory.

Based on these considerations, a valid opinion expression should: 1) describe subjectivity information; 2) contain at least a noun with at least a verb or with at least an adjective, optionally with adverbs as their modifiers, optionally with *Orientation* values (either *‘Positiv’* or *‘Negativ’*), and optionally with ‘Negate’ to indicate a ‘marked’ *Polarity*.

To satisfy the first condition, we restrict opinion expression extraction only in subjective sentences. Our method to extract opinion expressions is based on the *Apriori*, a frequent itemsets mining (or association rule mining) algorithm. Transaction records are generated from every sentence. Before describing the actual extraction method, we first discuss how a sentence is assigned an optional *Orientation* value and an optional ‘marked’ *Polarity* value.

To determine *Orientation* value, we adopt a strict method. A sentence is ‘Positiv’ if it contains at least a word that belongs to ‘Positiv’ category in GI and it contains no word that belongs to ‘Negativ’ at the same time. Similarly, a sentence is ‘Negativ’ if it contains at least a ‘Negativ’ word but without a ‘Positiv’ one at the same time. If a sentence does not contain any ‘Positiv’ or ‘Negativ’ word, or it contains both ‘Positiv’ word(s) and ‘Negativ’ word(s), no *Orientation* value is assigned to it. In this way, the assignment of *Orientation* value to a sentence will have a

[7] In GI, there are 1915 ‘Positiv’ words and 2291 ‘Negativ’ words.

[8] Apparently, *not* is the most common ‘Negate’ word. In addition to *not*, there are other words that can refer to reversal or negation, for example, *disapprove*. In GI, there are 217 such words.

TABLE III

TOP 10 OPINION EXPRESSIONS IN THREE GROUPS: 'LIBERAL ONLY', 'COMMON' AND 'CONSERVATIVE ONLY'

Liberal only	Common	Conservative only
american daily	conservative thompson	flag Negativ vote burn
liberal daily	aclu stop	american malkin michelle
progressive media	global democrat	conservative freedom
mad kane	democratic blue	muslim truth
political report politics	islamic america	george god link
political war politics	illegal alien Negativ	conservative michael
political wire	political correctness	american mike
progressive world	global al gore	senator flag vote burn
political daily kos	islamic muslim	conservative technorati
american kos	political leftist	muslim leftist

high accuracy with a low recall.

To resolve the *Polarity* value of a sentence, we accumulate all occurrences of words that belong to 'Negate' category in GI. A sentence has 'Negate' as a 'marked' *Polarity* value if it has totally odd number of occurrences of words that belong to 'Negate' category in GI.

Now, we describe our method to extract a list of opinion expressions from liberal blogs and a list of opinion expressions from conservative blogs. This means subjective sentences in liberal blogs and in conservative ones are separately processed. In the following, we will take liberal blogs as an example in our discussion. First, each sentence that is classified as *subjective* from each liberal blog is read, in which each word has its base form and is assigned with a POS tag. Second, for each POS-tagged word, look it up in GI to see if it belongs to 'Positiv' or 'Negativ', and/or 'Negate' and record the categories with the word if found. Next, determine the value of *Polarity* and the value of *Orientation* of each sentence. Then, for each sentence, list all nouns, verbs, adjectives, adverbs, 'Positiv', 'Negativ' and 'Negate' as a transaction record. So, each word (with its POS tagged), a 'Positive', a 'Negativ' and a 'Negate' is an item. Multiple occurrences of the same word with the same POS is listed only once. After all sentences are processed, we apply the *Apriori* algorithm to find maximum itemsets with the following settings: 1) The minimum support threshold is set to a value so that a frequent item set should appear at least once in every 10 blogs. 2) To avoid many irrelevant phrases are to be detected, we also set a maximum support threshold so that multiple items that co-occur at least in one sentence per blog are discarded. 3) Each maximum itemset should contain no more than five items. Lastly, among the maximum itemsets found, discard those that do not satisfy the definition of opinion expressions aforementioned. As a result, each of those remaining maximum itemsets is deemed as an opinion expression from liberal blogs. Following the same steps, we can obtain a list of opinion expressions from conservative blogs. Clearly, some opinion expressions will only appear in the list generated from liberal blogs, some will only appear in the list from conservative blogs, while some will appear in both lists. Therefore, we merge the two lists by joining on the same opinion expressions and then divide the merged list into the following three groups: 1) 'liberal only' opinion expressions that only appear in liberal blogs, 2) 'conservative only' opinion expressions that only appear in conservative blogs, and 3) 'common' opinion expressions that appear in both liberal and conservative blogs (note: though, the name 'common' does not necessarily mean all liberal and conservative bloggers hold the same opinions by these opinion expressions). The first two groups are ranked in descending order of frequencies of opinion expressions. To sort the third group, we first calculated odds ratio value of each opinion expression appearing more probably in liberal blogs and odds ratio value of it appearing more probably in conservative ones. Then, we use the summation of the two values to rank all the opinion expressions in the group in descending order.

Table III lists top 10 'liberal only', top 10 'conservative only' and top 10 'common' opinion expressions thus found. First, we can see that some of them point us to clear political events and opinions. One such example is a top 'conservative only' opinion expression, which is *flag Negativ vote burn*. In combination, the three words 'burn', 'flag' and 'vote' refer to the event of 'Flag-burning amendment fails by a vote'. And 'Negativ' may indicate that the majority attitude towards the event from the authors of the conservative blogs in our corpus is negative. Though, it can also be seen from the table that some are not opinion expressions and some do not provide clear clues as to what opinions are hold by bloggers on what issues.

The cause of such a mixture of interesting opinion expressions and non-opinion expressions in top results is due to the fact that certain number of irrelevant collocations of words are found by the *Apriori* algorithm. As a result, meaningful opinion expressions are scattered in all the three groups and are intervened by irrelevant collocations. So, a more reasonable way to use all the opinion expressions found is to issue meaningful queries by using nouns or noun phrases as indicators of issues or entities to find opinions on them, or by using verbs or adjectives, or 'Positiv', 'Negativ' and/or 'Negate' to find issues or entities "influenced" by them.

For example, as we know "immigration" is a popular political issue. To find out bloggers' opinions on this issue, we can use 'immigr' ('immigration' or 'immigrant' stemmed by Porter stemmer) as a query. One opinion expression found is ‘illegal immigration Negativ’, which appears 181 times in conservative blogs but only 55 times in liberal ones. Another result is ‘illegal driver immigrant’, which appears 66 times in conservative blogs, compared to 13 times in liberal ones. By issuing a query ‘cut’, we see an opinion expression ‘cut tax’, which is almost equally frequently used by both liberal and conservative bloggers (93 times vs. 98 times).

TABLE IV
TOP FIVE ‘POSITIV’, ‘NEGATIV’ AND ‘NEGATE’ OPINION EXPRESSIONS

Liberal	Conservative
gay marriage Positiv	real estate Positiv
democratic candidate Positiv	public school Positiv
democratic party Positiv	federal government Positiv
national security Positiv	site web page Positiv link
political party Positiv	bookmarking page Positiv discover link
american iraq Negativ	flag Negativ vote burn
military iraq Negativ	senator Negativ vote burn
human rights Negativ	illegal alien Negativ
foreign policy Negativ	0 trackback Negativ
iraq Negativ kill	illegal immigration Negativ
bush Negate	political bush Negate
war iraq Negate	american truth war Negate politically
administration Negate	click Negate
house Negate	war justice Negate
health Negate	bush democrat Negate

More interestingly, Table IV shows top five results of searching for ‘Positiv’, ‘Negativ’ and ‘Negate’, grouped by ‘liberal’ (ranked by the value of odds ratio that an opinion expression appears more probably in liberal blogs than in conservative ones) and ‘conservative’ (ranked by odds ratio that an opinion expression appears more probably in conservative blogs than in liberal ones).

From above search-based results, our method to extract opinion expressions demonstrates some power to reveal both the same and different opinions of liberal and conservative bloggers. To further verify this, we combined the opinion expression extraction process with 10-fold cross-validation in building political leaning classifiers to examine the performance. That is, each round the training set is used to generate opinion expressions. Then, these opinion expressions are used as binary features in building and testing classifier models - each blog is examined to set a value for each opinion expression feature (OE) : if at least a sentence contains the opinion expression, the corresponding feature value is 1; otherwise, the feature value is 0. Table V shows the results by using the combination of BoW and OE from all subjective sentences detected using GI, denoted by ‘$GIsubj_{BoW} + GIsubj_{OE}$’ and results by using BoW from all texts and OE from subjective sentences detected by GI, denoted by ‘$all_{BoW} + GIsubj_{OE}$’. For convenient comparison, results of both $GIsubj_{BoW}$ and all_{BoW} are also listed in the table, copied from Table II. However, to save space, only p-value of overall accuracy is shown for each classifier that uses opinion expression features (BoW+OE) compared to the one without using opinion expression features (BoW). It can be seen that by including opinion expressions, both classifiers achieved statistically significant improvement.

5 Discussion

First, using a rule-based subjective sentence classifier with the help of the General Inquirer dictionary shows success in improving political leaning classification. This also provides an extrinsic evaluation of the subjective classifier. Though, we plan to directly evaluate its performance in the future. Next, by only looking at subjective sentences, we also investigated top ‘discriminating words’ by calculating the odds ratio of each word appearing more probably in liberal blogs and the odds ratio of the word appearing more probably in conservative ones. However, these words do not provide a clear clue to reveal different opinions hold by bloggers belonging to the two opposite political orientations. This seems to confirm a previous observation we had that liberal and conservative blogs share the similar distribution of 182 GI semantic categories by only looking at individual words [33]. As a result, we need features that are beyond single words in order to find different opinions expressed in liberal and conservative blogs. Apparently, opinion expressions meet this requirement. By extracting opinion expressions, the same and different opinions expressed in liberal and conservative blogs become visible to us with the aid of formulating interesting queries to search for relevant opinion expressions. In addition, by including these opinion expressions as features, the performance of political leaning classification gets improved, demonstrating the effectiveness of the extraction method of the opinion expressions.

However, by examining Table III and Table IV, we see some are not opinion expressions and some are irrelevant to politics. This shows some weakness of the method that we used to extract opinion expressions. Though, this weakness is expected since the *Apriori* algorithm is designed to favor frequent itemsets, which are frequent co-occurrences of words and values of *Orientation* and *Polarity* in blogs. As a result, on one hand, the algorithm will inevitably find collocations of words that occur a lot but are irrelevant to opinions, which reduces the accuracy of finding valid opinion expressions. On the other hand, the algorithm favors popular issues and topics in blogs since they are talked a lot more than others, which brings about loss in recall. In the future, we need to design more effective methods to extract opinion expressions to remedy these drawbacks.

6 Conclusion and future work

In this paper, we have discussed the political leaning categorization problem and investigated the effectiveness of identifying and using subjective sentences in training SVM classifiers. Experiments showed an increase in categorization performance by using subjective sentences based on the General Inquirer dictionary. Especially, extracting collocations of nouns, verbs, adjectives and adverbs that meet certain criteria in a sentence as opinion expressions can enable us to see

TABLE V

POLITICAL LEANING CLASSIFICATION: BOW+OE VS. BOW

Classifier	L_P	L_R	L_F1	R_P	R_R	R_F1	Accuracy	Avg. F1
$GIsubj_{BoW} + GIsubj_{OE}$	89.34%	83.96%	0.8657	80.69%	86.27%	0.8339	84.96% ($p < 0.001$)	0.8498
$GIsubj_{BoW}$	87.65%	82.63%	0.8507	78.86%	84.13%	0.8141	83.28%	0.8324
$all_{BoW} + GIsubj_{OE}$	87.58%	82.82%	0.8513	78.89%	84.13%	0.8142	83.38% ($p < 0.0005$)	0.8328
all_{BoW}	86.01%	81.87%	0.8389	77.50%	81.98%	0.7968	81.92%	0.8179

both the same and different opinions hold by liberal and conservative bloggers on certain issues to some extent, and can enhance classification performance as well.

Our future work will be based on the following considerations. Firstly, we will design a better disambiguation method for GI semantic category lookup. We will also look for a more effective yet efficient method to extract opinion expressions to replace the current *Apriori*-based method. Lastly, we will consider to construct a domain knowledge base that contains a reasonable number of common issues that both liberal and conservative bloggers have interests. Clearly, this domain knowledge will provide guidance to extracting opinion expressions that plays the crucial role in political leaning categorization.

7 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*, 2005, pp. 36–43.

[2] 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*, 2006, pp. 159 – 162.

[3] K. T. Durant and M. D. Smith, "Mining sentiment classification from political web logs," in *Proceedings of WEBKDD'06*, 2006.

[4] D. Gibson, J. Kleinberg, and P. Raghavan., "Inferring web communities from link topology," in *Proceedings of HYPERTEXT '98*, 1998.

[5] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of web communities," in *Proceedings of KDD*, 2000, pp. 150–160.

[6] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of WWW*, 2007, pp. 221–230.

[7] 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.

[8] 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*, 2006.

[9] H.-J. Choi and M. S. Krishnamoorthy, "Categorization of blogs through similarity analysis," *Intelligence and Security Informatics, 2007 IEEE*, pp. 160–165, 2007.

[10] H. Koichi, F. Noriya, and T. Junichi, "Text categorization for Japanese blog entries," IPSJ SIG Technical Reports, Tech. Rep., 2005.

[11] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Mining the blogosphere: age, gender, and the varieties of self-expression," *First Monday*, vol. 12, no. 9, September 2007.

[12] M. Koppel, J. Schler, S. Argamon, and J. Pennebaker, "Effects of age and gender on blogging," in *Proceedings of AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[13] 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*, 2006.

[14] 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*, 2006.

[15] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proceedings of AIRWeb'05*, 2005.

[16] C. Fellbaum, *WordNet: an electronic lexical database*. MIT Press, 1998.

[17] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Comput. Linguist.*, vol. 30, no. 3, 2004.

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002, pp. 79–86.

[19] K. Bloom, S. Stein, and S. Argamon, "Appraisal extraction for news opinion analysis at NTCIR-6," in *Proceedings of the NTCIR 6*, 2007.

[20] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 417–424.

[21] E. Riloff, J. Wiebe, and W. Phillips, "Exploiting subjectivity classification to improve information extraction," in *Proc. 20th National Conference on Artificial Intelligence*, 2005, pp. 1106–1111.

[22] E. Spertus, "Smokey: automatic recognition of hostile messages," in *Proceedings of the Ninth Conference on Innovative Application of Artificial Intelligence, IAAI-97*, 1997, pp. 1058–1065.

[23] K. Eguchi and V. Lavrenko, "Sentiment retrieval using generative models," in *Proceedings of EMNLP'06*, 2006, pp. 345–354.

[24] L. Tari, P. H. Tu, B. Lumpkin, R. Leaman, G. Gonzalez, and C. Baral, "Passage relevancy through semantic relatedness," in *Proceedings of the Sixteenth Text REtrieval Conference, TREC 2007*, 2007.

[25] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of CICLing-2005*, 2005, pp. 486–497.

[26] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *ACM SIGIR CIKM'05*, 2005.

[27] J. R. Martin and P. R. White, *The Language of evaluation: the appraisal framework*. Palgrave Macmillan, 2005.

[28] M. A. Halliday and C. M. Matthiessen, *An introduction to functional grammar*. Arnold Publishers, 2004.

[29] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 105–112.

[30] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

[31] M. Jiang and S. Argamon, "Finding political blogs and their political leanings," in *Text Mining 2008, workshop at the SIAM International Conference on Data Mining*, April 2008.

[32] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of HLT/EMNLP 2005*, 2005, pp. 347–354.

[33] M. Jiang and S. Argamon, "Preliminary semantic analysis of political blogs," in *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM'08)*, 2008.