

Exploiting Subjectivity Analysis in Blogs to Improve Political Leaning Categorization

Maojin Jiang and Shlomo Argamon

Linguistic Cognition Lab, Department of Computer Science, Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616 USA
{jianmao, argamon}@iit.edu

ABSTRACT

In this paper, we address a relatively new and interesting text categorization problem: classify a political blog as either *liberal* or *conservative*, based on its political leaning. Our subjectivity analysis based method is twofold: 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 other features to build political leaning classifiers. Experimental results 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 leaning to some extent.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*linguistic processing*

General Terms

Algorithms, Experimentation

Keywords

political leaning categorization, subjectivity analysis

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 bloggers' different opinions on common issues that mark a border between liberal blogs and conservative ones. This makes the political leaning categorization both different from 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 in building political leaning classifiers. The work reported here is the extension of our previous one on the same problem, which used all texts in liberal and conservative blogs with 'Bag of Words' (BoW) and some link features [1]. In the following, we will first describe our corpus, and then discuss our methods and experiments.

2. CORPUS, CLASSIFIER LEARNING AND EVALUATION METRICS

Our corpus contains front pages of 1,054 liberal blogs and 793 conservative blogs, totalling more than 200MB. URLs of these blogs were obtained from five blog catalog web sites. All these blogs are written in English. In addition, the corpus contains no duplicate blog and no blog is labeled as both liberal and conservative.

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 SVM¹. All results reported in this paper were obtained by using SVM^{light} with linear kernels in a 10-fold cross-validation by setting 'cost-factor' to 0.75 ($\approx \frac{793}{1054}$) to balance the cost of 'false positive'² error and the cost of 'false negative'³ error.

We use the standard *precision* and *recall* to evaluate prediction of each category. We also calculate *F1* measure as follows: $F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. In addition, we calculate an overall accuracy as the number of correctly classified blogs of both categories divided by the total number of blogs to be classified. Lastly, average *F1* measure is 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. IDENTIFYING SUBJECTIVE SENTENCES

We use a rule-based subjective classifier to separate subjective sentences from objective ones. The rule of the 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 dictionary⁴ (GI) due to its abundant semantic categories. For our purpose, we chose the following nine categories, including, "Strong", "Hostile", "Pleasant", "Pain", "Feel", "Arousal",

¹<http://svmlight.joachims.org/>

²Conservative blogs are classified as 'liberal'.

³Liberal blogs are classified as 'conservative'.

⁴<http://www.wjh.harvard.edu/~inquirer/>

Table 1: Political Leaning Classifier Performance

Classifier	Size (%)	LP	LR	LF1	RP	RR	RF1	Accuracy	Average F_1
all_{BoW}	100	86.01%	81.87%	0.8389	77.50%	81.98%	0.7968	81.92%	0.8179
$GIsub_{BoW}$	63	87.65%	82.63%	0.8507	78.86%	84.13%	0.8141	83.28%	0.8324
$Glob_{BoW}$	37	75.54%	70.01%	0.7267	63.79%	69.85%	0.6669	69.94%	0.6968
$rand63_{BoW}$	63	81.66%	74.95%	0.7816	70.37%	77.31%	0.7368	75.96%	0.7592
$GIsub_{BoW} + GIsub_{OE}$	63	89.34%	83.96%	0.8657	80.69%	86.27%	0.8339	84.96%	0.8498

“EMOT”, “Virtue” and “Vice”. Thus, all words belonging to any of these nine categories are considered as strong subjective clues. In total, there are 4004 different entries (an entry is a word sense) in GI that appear on our strong subjective clue list.

The subjective classifier works as follows. First, we extract only text from the HTML file of a blog. Then the text is split into sentences after which part-of-speech (POS) tagging is applied to each sentence and each word is reduced to its base form. Next, each POS-tagged word is assigned to all categories of the same word with the same POS in GI if found. If a word belongs to at least one of the nine categories as mentioned, it is deemed as a *strong subjective clue*. Last, if a sentence contains at least two such clue words, it is classified as *subjective*. Otherwise, it is *objective*.

By applying the subjective classifier to both liberal and conservative blogs in our corpus, around 63% sentences in total were classified as *subjective*. To evaluate the subjective classifier, we compared results of a political leaning classifier by using BoW from these subjective sentences to those without exclusively using them. Table 1 shows the results, where ‘ all_{BoW} ’ denotes a classifier using all texts, ‘ $GIsub_{BoW}$ ’ is the classifier using only subjective sentences, ‘ $Glob_{BoW}$ ’ using only objective ones, and ‘ $rand63_{BoW}$ ’ using 63% of all sentences that are randomly selected. In addition, the column ‘Size’ indicates percentage of all sentences used in experiments. We can see that ‘ $GIsub_{BoW}$ ’ achieves the best performance among these classifiers and it outperforms ‘ all_{BoW} ’ on overall accuracy by 1.66% ($p < 0.005$).

4. EXTRACTING OPINION EXPRESSIONS

In our work, 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 category ‘*Positiv*’ or ‘*Negativ*’ to indicate an *Orientation* value, and optionally with ‘*Negate*’ to indicate a ‘marked’ *Polarity*. ‘*Positiv*’, ‘*Negativ*’ and ‘*Negate*’ are three GI categories. To meet the first requirement, we restrict opinion expression extraction only in subjective sentences. A sentence is ‘*Positiv*’ if it contains at least a ‘*Positiv*’ word but without a ‘*Negativ*’ one. Similarly, a sentence with at least a ‘*Negativ*’ word but with no ‘*Positiv*’ one is labeled ‘*Negativ*’. In other cases, no *Orientation* label is assigned. We also assign ‘*Negate*’ to a sentence if it contains odd number of occurrences of words in ‘*Negate*’ category in GI.

To extract opinion expressions, subjective sentences in liberal blogs and in conservative ones are separately processed. For each sentence, list all nouns, verbs, adjectives, adverbs, ‘*Positiv*’, ‘*Negativ*’ and ‘*Negate*’ as a transaction record. After all sentences are processed, we apply *Apriori* algorithm to find maximum itemsets with the following settings: 1) the minimum support threshold is set to such a value that a frequent itemset should appear at least once in every 10 blogs;

2) the maximum support threshold is set so that frequent itemsets appearing at least once per blog are discarded; 3) each maximum itemset should contain less than six items. Last, among all the maximum itemsets found, those meeting the above conditions are kept as opinion expressions. In this way, two lists of opinion expressions are obtained, one from liberal blogs and the other from conservative ones. Then the two lists are merged into one by joining on the same opinion expressions.

On the list of opinion expressions, we can query with nouns on interesting political issues to find opinions on them; we can also query with verbs to find political issues or entities “influenced” by these verbs. For example, by searching for ‘immigr’ (‘immigration’ or ‘immigrant’ stemmed by Porter stemmer), we obtained ‘illegal immigration Negativ’ (181 times in conservative blogs but 55 times in liberal ones). By searching for ‘cut’, we found ‘cut tax’, which was equally frequently used by both liberal and conservative bloggers. More interestingly, Table 2 shows top five ‘*Positiv*’, ‘*Negativ*’ and ‘*Negate*’ opinion expressions, grouped by ‘liberal’ and ‘conservative’.

Table 2: Top Five ‘*Positiv*’, ‘*Negativ*’ and ‘*Negate*’ Opinion Expressions

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

Lastly, by combining the opinion expression extraction process with a 10-fold cross-validation experiment, the last row in Table 1 shows the political leaning classification results by using BoW and using opinion expressions as binary features (OE), denoted by ‘ $GIsub_{BoW} + GIsub_{OE}$ ’. By including opinion expressions, the classifier achieved a statistically significant improvement of 2% over ‘ $GIsub_{BoW}$ ’ on overall accuracy ($p < 0.001$).

5. REFERENCES

[1] 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.