

Preliminary Semantic Analysis of Political Blogs

Maojin Jiang and Shlomo Argamon

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

Abstract

In this paper, we present a series of semantic analyses of words in political blogs in the setting of categorization of two opposite political orientations: liberal vs. conservative. We classify nouns, verbs, adjectives and adverbs into semantic categories by using the General Inquirer dictionary. Then distributions of these categories and correlations among them are examined both within and between blogs of the two opposite political leanings. Results show that although words of certain categories tend to appear together while others do not within blogs of a political leaning, the semantic category distribution of words used by left-wing bloggers is very similar to those by right-wing bloggers, suggesting single words alone do not account for major difference between these two major categories of blogs. Lastly, by examining preliminary results of association rule mining of nouns, verbs, adjectives and adverbs in sentences, we posit that the similarity and/or difference between blogs of opposite political orientations can be detected by extracting opinion expressions around collocation of nouns and verbs (together with modifiers).

Introduction

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. Though text content analysis and text categorization have been heavily researched, previous work on political blog categorization has been very limited. In (Durant and Smith 2006; Malouf and Mullen 2007), authors aimed to classify certain blog postings into opposite political orientations by mainly using “Bag of Words” features in their algorithms. But no discussion on where or when these features helped or failed from text analysis point of view was provided. In (Adamic and Glance 2005), authors used a phrase extraction algorithm to extract certain number of phrases as a phrase space and found intra-political-leaning blog similarity is higher than inter-political-leaning blog similarity by treating blogs as points in the phrase space.

Intuitively, it is bloggers’ different opinions on various issues that mark a border between liberal blogs and conservative ones. This has motivated us to analyze texts of political

blogs by grouping “content” words (in terms of bearing semantic meanings) based on their semantic meanings.

In the following, we will first describe our corpus, followed by discussion on our methods and experiments.

A Pilot Corpus

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 appears as both liberal and conservative.

Semantic Analysis

We chose the General Inquirer dictionary¹ (GI) due to its abundant categories. In total, it has 182 categories in 26 groups, mainly of nouns, verbs, adjectives and adverbs.

To assign a category to a word in a blog, we first extracted only text from HTML file of the blog. Then the text was split into sentences. Next, part-of-speech (POS) tagging was applied to each sentence and each word was reduced to its base form. Lastly, each POS-tagged word was assigned to all categories of the same word with the same POS in GI if found.

Then, we grouped all words by semantic categories in each blog, counted occurrence of each category and calculated its percentage in the blog. In this way, we obtained 1054 samples of distribution of 182 categories for liberal blogs and 793 samples for conservative blogs. This enables us to examine frequencies of semantic categories and their relations.

The result shows both liberal and conservative blogs have higher occurrences of “Positive” words (> 4.42%) than “Negative” ones (> 3.86%). Bloggers tend to use more “Strong” words (> 8.02%) than “Weak” words (< 2.50%). It is no surprise that both liberal and conservative bloggers write mostly on “POLIT” (> 5%), next on “ECON” (> 3%) and then on “Legal”, “Exprsv”, “Milit”, “Relig” and “Exch”.

In both liberal and conservative blogs, “POLIT” category has a positive correlation with these categories: “Strong”, “Power”, “ECON”, “Legal” and “Milit” while it has a negative correlation with “Weak”, “Pleasure”, “Arousal” and

¹<http://www.wjh.harvard.edu/~inquirer/>

Liberal	Active 9.23%	Strong 8.02%	IAV 7.14%	POWTot 7.13%	SV 5.85%	EnlTot 5.83%	POLIT 5.20%	HU 4.8%	Power 4.77%	Positive 4.42%
Conservative	Active 9.31%	Strong 8.07%	IAV 7.26%	POWTot 7.01%	EnlTot 5.98%	SV 5.96%	POLIT 5.09%	HU 4.76%	Power 4.74%	Positive 4.52%

Table 1: Top 10 most frequent semantic categories with average percentage of frequency per blog: liberal vs. conservative

“EMOT”. In addition, “POLIT” has a slightly higher positive correlation with “Positive” than with “Negative”. Surprisingly, we find that “Relig” category negatively correlates with “POLIT”, “Milit”, “Legal”, and “ECON”. This indicates that blog postings on religions are more focused than those on other topics. This may also indicate that there exist subcategories in either liberal or conservative blogs and some words in certain semantic categories can be discriminating in categorizing them. Similar results are seen between “Milit” and “Exch”, “Exprsv”, and “Academ” and are seen between “Academ” and “Legal”, “Exch”, and “Milit”.

Table 1 shows average percentage of frequency per blog of top 10 most occurring semantic categories in liberal and conservative blogs.

It can be seen that liberal and conservative blogs share the same top semantic categories with almost the same distribution. In fact, this is observed for almost all semantic categories between the two blog categories. By listing average percentage of a semantic category in each blog and calculating correlation between the list of all liberal blogs and the list of all conservative ones, it reveals as high as a 0.9999 positive correlation between the two distributions.

To further verify if semantic categories in liberal and conservative blogs share the same distribution, we did the following experiment. We first extracted and created separate lists of nouns, verbs, adjectives and adverbs in liberal blogs. Then we further divided each POS list into two: one containing words only appearing in liberal blogs and the other with words in both liberal and conservative blogs. Then we assigned semantic categories to words on these eight lists and gathered statistics. We processed the conservative blogs in a similar way. To our surprise, no noun, verb, adjective or adverb only appearing in either liberal or conservative blogs was assigned a semantic category label. We thus manually examined the lists and found most words are invalid English words due to free writing styles of bloggers (e.g. *super-wealthy*). So, this indicates either these words are not in GI or bear no semantic category. In all cases, less than 2% of such words appear more than 10 times in either liberal or conservative blogs. So, we think the inaccuracy caused by these words can be ignored.

For common word lists of nouns, verbs, adjectives and adverbs, we calculated odds ratio value of each word appearing more probably in liberal blogs and odds ratio value of it appearing more probably in conservative blogs. Next, we fetched top 7000 “liberal” nouns, verbs, adjectives and adverbs respectively and top 7000 “conservative” nouns, verbs, adjectives and adverbs respectively. Then, we calculated pair-wise correlation of semantic category distributions between top 7000 “liberal” and top 7000 “conservative” words of the same POS. Results show a very positive correlation

among the four pairs: 0.9127 for nouns, 0.9814 for verbs, 0.9693 for adjectives and 0.8045 for adverbs.

Lastly, we examined if semantic category distribution locally to a POS (noun, verb, adj or adv) differs from global distribution in all liberal blogs and in all conservative blogs. For this purpose, we took semantic category distributions of top 7000 “liberal” nouns, verbs, adjectives and adverbs (as above) and calculated correlation between each of them with global semantic distribution in liberal blogs. The correlation values are 0.6487, 0.7034, 0.3791 and 0.2845 for nouns, verbs, adjectives and adverbs respectively. This implies that semantic categories of blog text can be mostly predicted by verbs and nouns, then adjectives and adverbs.

Beyond Single Words

From above semantic analysis of blog text, it indicates that to fully distinguish a liberal blog from a conservative one, one must go beyond solely looking at single words. Since nouns and verbs account for major part of semantics in blogs, we believe the collocations of the nouns and verbs will be the foci of political blog categorization. In fact, this combination (together with modifiers) will form opinion expressions that carry opinions of bloggers on various issues.

To examine this idea, we extracted all nouns, verbs, adjectives and adverbs in each sentence as a transaction record. Next, we applied association rule mining tool to find association rules that contain at least a noun together with at least a verb or at least an adjective from all sentences in either liberal or conservative blogs. On the list of rules generated, we can query with noun or noun phrases on interesting political issues to find opinions on these issues; alternatively, we can query with verbs to find political issues or entities “influenced” by these verbs. By comparing rules from liberal blogs to conservative ones, it is possible to find out if liberal bloggers have different or similar opinions on common issues and this will help political blog categorization. In our experiment, for example, by searching for “border”, we found “border illegal->amnesty” and “enforce law->border” in conservative blogs but less intuitive results in liberal blogs. By searching for “cut”, we found “cut->tax” in both liberal and conservative blogs.

References

- Adamic, L. A., and Glance, N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *The 3rd International Workshop on Link Discovery*.
- Durant, K. T., and Smith, M. D. 2006. Mining Sentiment Classification from Political Web Logs. In *WEBKDD'06*.
- Malouf, R., and Mullen, T. 2007. Graph-Based User Classification for Informal Online Political Discourse. In *The 1st Workshop on Information Credibility on the Web*.