# Graph-based user classification for informal online political discourse

Robert Malouf[1] and Tony Mullen[2]

[1] Department of Linguistics and Asian/Middle Eastern Languages
San Diego State University
rmalouf@mail.sdsu.edu
[2] Department of Computer Science
Tsuda College, Tokyo
mullen@tsuda.ac.jp

**Abstract.** With the rise of the interactive "Web 2.0" and the increasing tendency of online publications to turn to message-board style reader feedback venues, informal political discourse has become an important feature of the intellectual landscape of the Internet. We consider innate political bias or "sentiment" to be of interest for a variety of reasons, including as a factor in determining the reliability of posters in terms of authority and truthfulness. We describe several experiments in identifying the political orientation of posters in an informal environment. Our results indicate that the most promising approach is to augment text classification methods by exploiting information about how posters interact with each other.

## 1 Introduction

The rise of the interactive 'Web 2.0' is changing the nature of typical web texts and has raised significant new challenges for natural language processing. Until recently, much of the text available on the WWW was either professionally edited or followed the conventions of edited text. Newspapers, magazines, corporate and government publications, and academic papers, and even personal and hobby sites produced by amateurs follow fairly rigid standards for formatting, style, and orthography. More importantly, they are intended to be read by a large anonymous audience which shares only the most general public context. This makes them easy to process with relatively little specific background knowledge, both for human readers who may be referred to the site by a search engine, and for automated methods such as question answering or text mining systems.

On the other hand, in environments such as discussion forums, social networking sites, and chat rooms, where content is submitted by the users themselves, the use of language is very different. Unlike edited text, informal web texts are typically conversational, are often non-standard or idiosyncratic, and are highly contextualized, depending on rich background of shared knowledge and assumptions.

Informal web texts pose new and interesting problems for text processing techniques which have been developed for more traditional edited text genres. In this paper we will explore methods for *sentiment analysis* in informal political texts. Sentiment analysis refers to the task of identifying opinions, favorability judgments, and other information

related to the feelings and attitudes expressed in natural language texts. Our research investigates the application of similar techniques to the political domain, in particular the domain of informal political discourse. Political sentiment can be useful in a variety of ways, both as identifying the mindset of a potential audience of posters and as a means of recognizing underlying ideological biases that could have an impact on how reliable a source of information is assumed to be.

While some work has been done on sentiment analysis for political texts [1, 2], the extent to which this task differs from more conventional sentiment analysis tasks has not been fully explored. In this paper we follow work reported in [3] using a data set of political discourse data from an online American politics discussion group.

## 2 Analysis of politically relevant sentiment

There are many applications for recognizing politically-oriented sentiment in texts. These applications include analyzing political trends within the context of a given natural language domain as a means of augmenting opinion polling data; classifying individual texts and users in order to target advertising and communications such as notices, donation requests or petitions; and identifying political bias in texts, particularly in news texts or other purportedly unbiased texts. This last use is particularly pertinent to evaluating the reliability of information sources, since it is widely assumed that an excess of political bias is a corrupting factor on the reliability of an information source.

Many of the challenges of the present task are analogous, but not always identical, to those faced by traditional sentiment analysis. It is well-known that people express their feelings and opinions in oblique ways. Furthermore, unlike opinion as addressed in conventional sentiment analysis, which focuses on favorability measurements toward specific entities, political attitudes generally encompass a variety of favorability judgments toward many different entities and issues.

### 2.1 Challenges in processing the data

The data we analyze has two distinct defining characteristics: its predominantly political content and its informality. Each of these qualities introduces challenges and methods of addressing these challenges can sometimes interfere with each other. One of the difficulties with analysis of informal text is dealing with the considerable problem of rampant spelling errors. This problem is compounded when the work is in a domain such as politics, where jargon, names, and other non-dictionary words are standard. The domain of "informal politics" introduces jargon all of its own, incorporating terms of abuse, pointed respellings, and domain specific slang. The difficulties of analysis on the word level percolate to the level of part-of-speech tagging and upwards, making any linguistic analysis challenging.

### 2.2 Political sentiment analysis as a classification task

For the present task, we conducted tests using several classification schemes. We used both the hand-modified self-descriptions as they stood, and we used a more general

classification of *right*, *left*, and *other*, which was composed of people who described themselves as "centrist", "libertarian" or "independent." The hand-modification we did on the self-descriptions was usually straightforward, although in one instance a self-described "Conservative Democrat" was modified to "conservative." If there had been enough conservative Democrats in the data to justify it, this classification probably should have been allowed to stand as a distinct self-described class, and generalized to the *other* class.

## 3 Data resources

### 3.1 The politics.com discussion database

We created a database of political discourse downloaded from `www.politics.com`. The database consists of approximately 77,854 posts organized into topic threads, chronologically ordered, and identified according to author, author's stated political affiliation. Furthermore, the posts are broken down into smaller chunks of text based on typographical cues such as new lines, quotes, boldface, and italics, which represent segments of text which may be quotes from other authors. Each text chunk of three words or greater is identified as quoted text or non-quoted text based upon whether it is identical to a substring in a previous post by another poster. The database contains 229,482 individual text chunks, about 10 percent of which (22,391 chunks) are quotes from other posts.

The total number of individual posters is 408. The number of posts by each author follows an inverse power-law distribution, with 77 posters (19%) logging only a single post. The greatest number of posts logged by a single poster is 6,885 posts, followed by the second greatest number of posts at 3,801 posts.

| | | | |
|---|---|---|---|
| RIGHT | 34% | Republican | 53 |
| | | Conservative | 30 |
| | | R-fringe | 5 |
| LEFT | 37% | Democrat | 62 |
| | | Liberal | 28 |
| | | L-fringe | 6 |
| OTHER | 28% | Centrist | 7 |
| | | Independent | 33 |
| | | Libertarian | 22 |
| | | Green | 11 |
| | | Unknown | 151 |

**Fig. 1.** Distribution of posts in the data by general class and by a slightly modified version of the writers' own self-descriptions.

### 3.2 Other data

In addition to the main dataset used for training and testing, additional data from the web was used to support spelling-correction. For this, we used 6481 politically oriented syndicated columns published online on right and left leaning websites `www.townhall.com` and `www.workingforchange.com` (4496 articles and 1985 articles, respectively). We also used a wordlist of email, chat and text message slang, including such terms as "lol," meaning "laugh out loud."

## 4 Sentiment analysis

To test the applicability of sentiment analysis methods to predicting user's political affiliation, we applied a variant of Turney's [4] *PMI-IR* method. In Turney's original application, opinion oriented texts such as reviews are tagged and descriptive phrases are extracted according to POS based templates. A value for the "semantic orientation" (SO) of each phrase is identified by find occurrences of target phrase near either the word "excellent" or the word "poor" in a reference corpus and pointwise mutual information (PMI) of each phrase with "excellent" and "poor" is calculated. Pointwise mutual information (PMI) is an information theoretic measure of how much two events tend to occur together For two events $x$ and $y$:

$$\mathrm{PMI}(x,y) = \log \frac{P(x,y)}{P(x)\,P(y)}$$

The semantic orientation (SO) of a term is:

$$\mathrm{SO}(w) = \mathrm{PMI}(w, \text{excellent}) - \mathrm{PMI}(w, \text{poor})$$

The overall orientation of a text is considered to be the average of the SOs of the phrases in the text. Turney's method yielded overall accuracy of 74.39%, although the results varied widely across domains (60%–85%).

In principle, the same method could be used for any one-dimensional classification (liberal vs. conservative, cheap vs. expensive, etc.) We measure political SO using PMI with the terms "liberal" and "conservative":

$$\mathrm{SO}(w) = \mathrm{PMI}(w, \text{liberal}) - \mathrm{PMI}(w, \text{conservative})$$

PMI scores for words were computed based on counts from the 200 million word Reuters news corpus RCV1 [5]. We extracted 171,617 noun phrases using a tag sequence filter. Of these, 5,183 occur in context with liberal or conservative more than three times in the reference corpus.

When viewing individual orientation values for phrases, the results were often (but not always) intuitive. Some examples can be seen in figure 2. We took a very naive approach to using this method for classification, along the lines of Turney by classifying texts for 184 users who can be classed as either LEFT (Democrat, liberal, L-fringe) or RIGHT (Republican, conservative, R-fringe), by averaging the orientation values of the phrases in their posts. This yielded a 40.76% accuracy in classifing posters. Informal

| | | | |
|---|---|---|---|
| jerry falwell | −6.160 | nuclear technology | 3.245 |
| bill kristol | −6.119 | euthanasia | 3.325 |
| social mores | −5.937 | health care system | 3.344 |
| weekly standard | −5.736 | lib dems | 4.423 |
| pat buchanan | −5.404 | condom use | 4.922 |
| judicial watch | −5.377 | ruth bader ginsburg | 5.238 |
| american enterprise institute | −5.290 | economic policy institute | 5.719 |
| far rightists | −5.244 | | |

**Fig. 2.** Some phrases and the values resulting from PMI-IR methodology applied with "conservative" and "liberal". Negative values indicate a stronger association with the word "conservative" than "liberal", and positive values indicate a stronger association with "liberal" than with "conservative".

manual annotation suggests a ceiling of 87.50%, whereas simply assigning the most frequent label (LEFT) yields a baseline of 52.17% accuracy.

Clearly, the results of applying this Turney-inspired method in the present manner to political texts are less than encouraging. There are some plausible explanations for why this approach performs so poorly. It could well be that the RCV1 corpus is simply too small to meet the demands of this task. Since the PMI-IR method is a form of unsupervised learning, a very large amount of data is generally necessary to get good results. However, the fact that PMI-IR's accuracy is so far *below* the baseline suggests that the political orientation of the noun phrases in these texts may be unrelated or even inversely related to the posters' own political orientations.

## 5 Text classification

To test the effectiveness of standard text classification methods for predicting political affiliation, we divided the users into the two general classes RIGHT (Republican, conservative, and r-fringe) and LEFT (Democrat, liberal, and l-fringe), setting aside the centrist, independent, green, and libertarian users. We then used the naive Bayes text classifier Rainbow [6] to predict the political affiliation of a user based on the user's posts. There were 96 users in the *left* category and 89 in the *right*, so a baseline classifier which assigned the category LEFT to every user would yield 52.17% accuracy. The NB text classifier gave an accuracy of 63.59% a modest (though statistically significant) improvement over the baseline.

There are a few possible explanations for the poor performance of a text classifier on this task. One hypothesis is that the language (or at least the words) used in political discussions does not identify the affiliation of the writer. For example, for the most part posters from across the political spectrum will refer to "gun control" or "abortion" or "welfare" or "tax cuts", regardless of their stance on this particular issues [1].

Another possibility is that irregular nature of the texts poses a special challenge to classifiers. The posts in the database are written in highly colloquial language, and are full of idiosyncratic formatting and spelling. Irregular spellings have a particularly harmful effect on lexically-based classifiers like Rainbow, greatly increasing the amount

of training data required. To test the contribution of users' misspellings to the overall performance, we ran all the posts through `aspell`, a freely available spell check program, augmented with the list of political words described in section 3.2. For each word flagged as misspelled, we replaced it with the first suggested spelling offered by `aspell`. Repeating the NB experiments using the corrected text for training and evaluation gave us an overall accuracy of 58.7%, significantly worse ($p = 0.03$) than the model without spelling correction.

A third possibility is that the disappointing performance of the classifier might be related to the skewed distribution of posting frequency. The corpus contains only a small amount of text for users who only posted once or twice, so any method which relies on textual evidence will likely have difficulty. There is some evidence that this is part of the problem. We repeated the NB experiments but restricted ourselves to frequent posters (users with more than a total of 500 words observed). With this restricted dataset, a baseline classifier gives 53.0%, and the human ceiling is 91.00%. Applying Naive Bayes to the subset of frequent posters yields 67.00% accuracy, again, a significant improvement over the baseline.

These results suggest two things. First, the performance of the classifier is very sensitive to the amount of training data used. And, second, any classifier will perform better for frequent posters than for light posters. Fortunately, simply collecting more posts will give us a large database to train from and will solve the first problem. However, it will not solve the second problem. Due to the 'scale free' nature of the distribution of posting frequency, any sample of posts, no matter how large, can be expected to include a substantial fraction of infrequent posters. In addition, even for frequent posters the results are somewhat disappointing.

Since purely text-based methods are unlikely to solve the problem of predicting political affiliations by themselves, we also looked at using the social properties of the community of posters. Unlike web pages, posts rarely contain links to other websites. However, many posts refer to other posts by quoting part of the post and then offering a response or by addressing another poster directly by name.

Of the 41,605 posts by users classified as either LEFT or RIGHT, 4,583 included quoted material from another user who could also be classified as either LEFT or RIGHT. Of these, users strongly tended to quote other users at the opposite end of the political spectrum. LEFT users quote RIGHT users 62.2% of the time, and RIGHT users quote LEFT users 77.5% of the time. In this respect the quoting relationship between posts appears to be markedly different from the inter-blog linking relationship discussed in Adamic and Glance [7], in which liberal and conservative blog sites are shown largely to link to other sites of agreeing political outlook.

To exploit this source of information, we used patterns of shared co-citations to group users into "teams". We first constructed a graph representing citation patterns, with each user represented by a node and each quoted post represented by an edge. We then computed a low-rank approximation (via singular value decomposition) of the citation graph's adjacency matrix, to reduce noise and to highlight second-order structural generalizations [8]. We then computed the distance between each pair of users

in the resulting 'citation space':

$$\text{dist}(\mathbf{x},\mathbf{y}) = 1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Using these distances, we clustered the users to find groups of posters with similar citation patterns. Since posters who share citation habits are playing a similar conversational role in the on-line political debates, we assume that all users in a cluster share a political affiliation. Therefore, we treated each cluster of users as if it were a single aggregate user. We applied the Naive Bayes classifier discussed above to all the posts in the cluster, and then assigned the predicted affiliation to all the users in the cluster.

This approach yielded more promising results. For all users, this approach yields 68.48% accuracy, a significant ($p = 0.003$) improvement over simple Naive Bayes. And, for users with >500 words, this improves to 73.00% with clustering, a significant ($p = 0.03$) improvement over Naive Bayes alone.

| Method | All users | >500 words |
|---|---|---|
| Baseline | 52.17 | 53.00 |
| Turney-inspired | 40.76 | 50.00 |
| NB | 63.46 | 67.00 |
| Cluster+NB | 68.48 | 73.00 |
| Human | 87.50 | 91.00 |

**Table 1.** Summary of results

## 6 Conclusions and future work

Our results suggest that information gained from the discourse relations between posters is of use in identifying the political sentiment of the posts.

A number of technical improvements could be made on the approaches we describe in this paper. SVM modeling could be used in place of Naive Bayes and better clustering algorithms may be of help. Further information within posts may be available to improve link detection, such as vocatives used when one poster is directly addressing another by name or handle. Group identifiers such as occur in the quote *Its very sad you conservatives cant win arguments on your own merit, you have to go petty and clone us liberals* could be exploited more fully. Thread structure itself may also yield clues to better construct teams and find links.

There is still much left to investigate in terms of optimizing the linguistic analysis, beginning with spelling correction and working up to shallow parsing and co-reference identification. Likewise, it will also be worthwhile to further investigate exploiting sentiment values of phrases and clauses, taking cues from methods such as those presented in Wilson, et al. [9], Nasukawa [10], and Turney [11]. Variants on the Turney method

may prove to be of use in identifying attitudes towards specific topics, which could then be used as features in a more general model.

## 7 Acknowledgments

Funding for this research was provided by a grant from the Japan Society for the Promotion of Science.

## References

1. Efron, M.: Cultural orientation: Classifying subjective documents by cociation analysis. In: AAAI Fall Symposium on Style and Meaning in Language, Art, and Music. (2004)
2. Efron, M., Zhang, J., Marchionini, G.: Implications of the recursive representation problem for automatic concept identification in on-line governmental information. In: Proceedings of the ASIST SIG-CR Workshop. (2003)
3. Mullen, T., Malouf, R.: A preliminary investigation into sentiment analysis of informal political discourse. In: Proceedings of the AAAI-2006 Spring Symposium on "Computational Approaches to Analyzing. (2006)
4. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, ACL (2002) 417–424
5. Lewis, D., Yang, Y., Rose, T., Li., F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5** (2004) 361–397
6. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow` (1996)
7. Adamic, L., Glance, N.: The political blogosphere and the 2004 u.s. election: Divided they blog. In: Proceedings of the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan (May 2005)
8. Drineas, P., Krishnamoorthy, M., Sofka, M., Yener, B.: Studying e-mail graphs for intelligence monitoring and analysis in the absence of semantic information. In: Intelligence and Security Informatics. Volume 3073 of Lecture Notes in Ccomputer Science. Springer-Verlag (2004) 297–306
9. Wilson, T., Wiebe, J., Hoffman, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT-EMNLP. (2005)
10. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. In: The Second International Conferences on Knowledge Capture (K-CAP 2003). (2003)
11. Turney, P.: Measuring semantic similarity by latent relational analysis. In: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland (2005) 1136–1141