A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse

Tony Mullen
Department of Computer Science
Tsuda College
Tsuda-machi 2-1-1, Kodaira-shi
Tokyo, Japan
mullen@tsuda.ac.jp

Robert Malouf
Department of Linguistics and Oriental Languages
San Diego State University
5500 Campanile Dr
San Diego, CA 92182-7727 USA
rmalouf@mail.sdsu.edu

Abstract

With the rise of weblogs and the increasing tendency of online publications to turn to message-board style reader feedback venues, informal political discourse is becoming an important feature of the intellectual landscape of the Internet, creating a challenging and worthwhile area for experimentation in techniques for sentiment analysis. We describe preliminary statistical tests on a new dataset of political discussion group postings which indicate that posts made in direct response to other posts in a thread have a strong tendency to represent an opposing political viewpoint to the original post. We conclude that traditional text classification methods will be inadequate to the task of sentiment analysis in this domain, and that progress is to be made by exploiting information about how posters interact with each other.

Introduction

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. The desirability of automatically identifying such information as it pertains to products, companies and other commercial entities is well established and the subject of considerable research (Turney & Littman 2003; Pang & Lee 2004; Morinaga, Kenji Yamanishi and, & Fukushima 2002; Mullen & Collier 2004). Sentiment analysis can be useful as a means of automatically handling customer feedback, as a basis for targeting advertising, and as a tool to assist in analyzing consumer trends and tendencies.

Our research seeks to investigate the application of similar techniques to the political domain, in particular the domain of informal political discourse. With the rise of weblogs and the increasing tendency of online publications to turn to message-board style reader feedback venues, informal political discourse is becoming an important feature of the intellectual landscape of the Internet.

While some work has been done on sentiment analysis for political texts (Efron 2004; Efron, Zhang, & Marchionini 2003), the extent to which this task differs from more conventional sentiment analysis tasks has not been fully explored. In this paper we introduce a new dataset of political discourse data from an online American politics discussion group. We report the results of a variety of statistical tests on the data to form a clear picture of the nature of the task, what it will entail and the most promising angles of approach to the particular problems it presents. We find that simple text classification methods will probably not yield very impressive results and that exploiting the interactive nature of the dialogue is likely to be the best way forward.

Motivation

As in the commercial domain, there are many applications for recognizing politically-oriented sentiment in texts. These applications include, among others, 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.[1]

Analysis of politically relevant sentiment

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. Word-based models succeed to a surprising extent but fall short in predictable ways when attempting to measure favorability toward entities. Pragmatic considerations, sarcasm, comparisons, rhetorical reversals ("I was *expecting* to love it"), and other rhetorical devices tend to undermine much of the direct relationship between the words used and the opinion expressed. Any task which seeks to extract human opinions and feelings from texts will have to reckon with these challenges. However, unlike opinion as addressed in conventional sentiment analysis, which focuses on favorability measurements toward specific entities, political attitudes generally encompass a

[1] It is worth commenting that methods of political sentiment analysis may also lend themselves to potentially abusive applications, such as use by unscrupulous or oppressive governments to censor or otherwise persecute dissent. While this is regrettable, the authors believe that responsibility for the protection of individual rights lies with an accountable and transparent government answerable to the rule of law.

variety of favorability judgments toward many different entities and issues. These favorability judgments often interact in unexpected or counterintuitive ways. In the domain of American politics, for example, it is likely that knowing a person's attitude toward abortion will help to inform a guess a that person's attitude toward the death penalty. Furthermore there are other political sentiment-related questions we may wish to ask about a text, aside from simply favorability judgments toward a specific issue, candidate, or proposal. These may include:

- Identifying the writer's political party affiliation
- Classifying the writer's political viewpoints according to some more general taxonomy, such right vs. left
- Gaging the "extremeness", or distance from a politically centrist position, of the writer's views
- Evaluating the degree of confidence with which the writer expresses her opinions
- Evaluating the degree of agreeability/argumentativeness with which the writer communicates
- Identifying particular issues of political importance to the writer

Challenges in processing the data

The data we wish to analyze has two distinct defining characteristics: its predominantly political content and its informality. Each of these qualities introduces particular challenges and methods of addressing these challenges can sometimes interfere with each other. One of the primary difficulties with analysis of informal text, for example, 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 (such as the spelling of "Reagan" as the homophone "Raygun" as a comment on the former president's support for the futuristic "Star Wars" missile defense project), and domain specific slang (such as "wingnuts" for conservatives and "moonbats" for liberals).

The difficulties of analysis on the word level percolate to the level of part-of-speech tagging and upwards, making any linguistic analysis challenging. For this reason, named-entity recognition, automatic spelling correction, and facility at handling unknown words would seem to be of crucial importance to this task. Even if this is accomplished, however, the lack of organization persists at higher levels. Grammar is haphazard, and rhetorical organization, to the extent that it is present at all, is unreliable.

Political sentiment analysis as a classification task

The first practical question which must be addressed is what specific information we are after and how to couch the task in terms of machine learning. We assume that we will approach the task as a classification task. So what are the classes?

There is an element of arbitrariness in any selection of classes we might make. Political sentiment, as suggested above, is not a simple binary classification. Although the traditional right/left distinction is an obvious possibility, it is not enough to describe the various shades of American political thought. Other taxonomies exist which take into consideration more information, such as attitudes toward the structure and influence of government, personal and economic freedom, rationality, and other factors. [2] It may not be necessary to model such nuances in practice, however. The classification scheme we decide on will need to reflect real divisions in the texts if it is to be modelable, but it will also depend largely upon practical considerations of what information we have decided we wish to extract.

A related issue in practice is that of the kind of information we have available as training data. In the current dataset of political discussion posts, class information for training is derived from the self-described political affiliation of the writers. Writers are given total freedom in their descriptions, and so some of the political affiliations were translated by hand into standard terms. A description such as "true blue" was translated to "democrat", whereas "USA Skins" was translated into "r-fringe." Using a combination of verbatim self-descriptions and hand-made general classes, we arrived at a classification including: *centrist*, *liberal*, *conservative*, *democrat*, *republican*, *green*, *libertarian*, *independent*, *l-fringe* and *r-fringe*. Obviously, there are overlaps here, and some distinctions may not be worth modeling. In terms of political attitudes, it is unlikely that a division between "liberal" and "democrat" is going to be useful in many applications. Nevertheless, from these classes it is already clear that a simple right/left distinction will leave some classes difficult to classify. Self-described Libertarians, centrists, and independents all create problems for a binary left-right classification scheme. Another question is whether the voices at the extremes are properly classified with moderates. Certainly the terminology used by a neo-Nazi skinhead who claims to worship Odin bears little in common with that of a small-government, fiscally conservative Republican, even if they are both classified as right of center. Even among members of particular political parties, views can be deeply divided (Pew Research Center 2005).

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.

[2] `http://en.wikipedia.org/wiki/Political_spectrum`

Right	24%	Republican	8%
		Conservative	16%
		R-fringe	0%
Left	46%	Democrat	35%
		Liberal	7%
		Green	1%
		L-fringe	3%
Other	30%	Centrist	2%
		Independent	11%
		Libertarian	17%

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

Data resources

The `www.politics.com` discussion database

We have 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 a Zipf-like distribution, with 77 posters (19%) logging only a single post. The greatest number of posts logged by a single poster is 6885 posts, followed by the second greatest number of posts at 3801 posts.

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."

Evaluation

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 (McCallum 1996) 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 51.89% accuracy. The NB text classifier gave an accuracy of 60.37% with a standard deviation of 2.21, based on 10-fold cross validation, While this is a statistically significant improvement over the baseline, it is modest.

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 (Efron 2004).

Another possibility is that irregular nature of the texts poses a special challenge to classifiers. Like all web text, 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, ran all the posts through `aspell`, a freely available spell check program, augmented with the list of political words described in section . 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 60.37% with a standard deviation of 1.12, which represented no improvement over 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 who posted twenty or more times). There were 50 frequent posters in each class, giving us a baseline of 50.0%. Since restricting the data this way reduces the number of training examples, we would expected to see the accuracy of the classifier to we reduced. And, if we train and evaluate a classifier on 50 randomly selected posters from each class, we get an accuracy of 52.00% which, with a s.d. of 3.27, is not significantly different from the baseline. However, when use posts from 100 frequent posters to train and evaluate the classifier, we get an accuracy of 61.38% (with a standard deviation of 1.60). With spelling correction, the result was 64.48% (2.76). It is possible that the spelling corrections yielded some improvement here, but more tests are needed to determine if the improvement is statistically significant. It is worth noting that the approach to spelling correction we use here is quite crude and results in many mis-corrected words. Some simple heuristics for spelling correction may go a long way toward improving the usefulness of this step.

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.

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 & Glance(2005), in which liberal and conservative blog sites are shown largely to link to other sites of agreeing political outlook. The pattern here suggests a simple classification rule: assign a user the opposite political affiliation of the users they tend to quote or be quoted by. If we assume that the affiliation of all quoted users is known, this rule yields 77.45% accuracy for those users who quoted at least one post or had at least one post quoted by another user. However, since this covers only 55.7% of the users, this rule has an overall accuracy of only 65.57%, still an improvement over the NB classifier.

Conclusions and future work

Our analysis of the data suggests that traditional word-based text classification methods will be inadequate to the task of political sentiment analysis. Some of the trouble may derive from the fact that in an argument about a given topic, both sides are likely to be using largely the same vocabulary. More generally, as Turney(2002) observed, sentiment analysis tasks become more difficult as the topic becomes more abstract. The language people use to describe their feelings about art, for example, tends to be less concrete than the language they use to evaluate a product. It is reasonable to speculate that the language used in political discourse lies on the more oblique end of this spectrum.

We expect that the most fruitful approach to this problem will be to incorporate a combination of models. We intend to look further into optimizing the linguistic analysis, beginning with spelling correction and working up to shallow parsing and co-reference identification. From there we intend to attempt a variety of approaches to evaluating sentiment values of phrases and clauses, taking cues from methods such as those presented in Wilson, Wiebe, & Hoffman (2005), Nasukawa & Yi (2003), and Turney (2005). In addition to using this information, we will attempt to exploit the discourse structure of the data, analyzing more fully how posters interact with each other in order socially and pragmatically, to use what we know about one poster to help classify others.

References

Adamic, L., and Glance, N. 2005. 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*.

Efron, M.; Zhang, J.; and Marchionini, G. 2003. Implications of the recursive representation problem for automatic concept identification in on-line governmental information. In *Proceedings of the ASIST SIG-CR Workshop*.

Efron, M. 2004. Cultural orientation: Classifying subjective documents by cociation analysis. In *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*.

McCallum, A. K. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow`.

Morinaga, S.; Kenji Yamanishi and, K. T.; and Fukushima, T. 2002. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 341 – 349.

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*.

Nasukawa, T., and Yi, J. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *The Second International Conferences on Knowledge Capture (K-CAP 2003)*.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 271–278. AC.

Pew Research Center. 2005. The 2005 political typology: Beyond red vs blue: Republicans divided about role of government - democrats by social and personal values. News Release.

Turney, P., and Littman, M. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21(4):315–346.

Turney, P. 2002. 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*, 417–424. Philadelphia, Pennsylvania: ACL.

Turney, P. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 1136–1141.

Wilson, T.; Wiebe, J.; and Hoffman, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*.