Taking sides: User classification for informal online political discourse

Robert Malouf
Department of Linguistics and Asian/Middle Eastern Languages
San Diego State University
San Diego, CA 92182-7727
USA
rmalouf@mail.sdsu.edu

Tony Mullen
Department of Computer Science
Tsuda College
2-1-1 Tsudamachi, Kodairashi
Tokyo 187-8577
JAPAN
mullen@tsuda.ac.jp

ABSTRACT

(Research paper)

Purpose

To evaluate and extend existing natural language processing techniques into the domain of informal online political discussions.

Design/methodology/approach

A database of postings from a U.S. political discussion site was collected, along with self-reported political orientation data for the users. A variety of sentiment analysis, text classification, and social network analysis methods were applied to the postings and evaluated against the users' self-descriptions.

Findings

Purely text-based methods performed poorly, but could be improved using techniques which took into account the users' position in the online community.

Research limitations

The techniques we applied here are fairly simple, and more sophisticated learning algorithms may yield better results for text-based classification.

Practical implications

This work suggests that social network analysis is an important tool for performing natural language processing tasks with informal web texts.

Originality/value

This research extends sentiment analysis to a new subject domain (U.S. politics) and a new text genre (informal online discusssions).

Taking sides: Graph-based user classification for informal online political discourse

1 Introduction

The development 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, 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 created 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. The only audience the text is intended for is the immediate participants in the online discussion at the time the text was produced.

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 aspects of 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. Analysis of 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.

Figures 1 and 2 illustrate the difference in how closely related opinions may be expressed in more and less formal language, depending on the nature of the text.

take in Figure 1

In Figure 1, we see an excerpt from a newspaper editorial on the minimum wage. [1] As readers, we know it is an editorial—a statement of institutional opinion written by the editors of a newspaper—because it is labeled as such, just as we know that the authors are the editors of *The Albuquerque Tribune,* a newspaper in a mid-sized city in the Southwestern United States. From the additional meta-data in the header, we also know when the article was first published, and the general opinion being expressed. The text of the article then lays out an argument

supporting the overall thesis, along with a set of supporting facts. With a minimum of background knowledge, the opinion expressed in this article can be understood and placed in context by the generic mass of newspaper readers who make up its intended audience.

take in Figure 2

In contrast, the opinions expressed in the examples in Figure 2 are much more difficult to process. These examples are taken from an on-going discussion on the minimum wage taken from the `politics.com` dataset (for details, see below). Each of these excerpts expresses an opinion about the minimum wage plus a supporting argument for that position, but, taken out of context, it is very difficult for a reader to evaluate these arguments or even to understand what opinion is being expressed. Unlike newspaper editorials, these texts were not produced for the benefit of an anonymous audience. Instead, they were produced as a by-product of an interaction, in which all that mattered as that the participants understood what was being expressed at the time.

Furthermore, even if we can process the meaning expressed by examples like these, it is very difficult to place them in a more global context. In contrast to signed newspaper articles or even unsigned editorials, we generally know almost nothing about the authors of informal web texts beyond their choice of screen name. The participants in online discussions could be professionals and academics or highly knowledgeable amateurs, or they could be schoolchildren. They could be expressing their own heartfelt beliefs, or they could be deliberately

taking an extreme position to incite outrage. We do not even know how many individual participants there are—one person might post under several screen names, or several people might post under one. Without even minimal knowledge about the author or the author's motivation, the reader has know way to evaluate the opinions being expressed.

2 Analysis of politically relevant sentiment

The desirability of automatically identifying an author's sentiment with respect to a topic 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. 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, Thomas, Pang & Lee 2006), the extent to which this task differs from more conventional sentiment analysis tasks has not been fully explored. In this paper we expand on our earlier work (Mullen & Malouf 2006, Malouf & Mullen 2007) using a data set of political discourse data from an online American politics discussion group.

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. Furthermore, expanding the domain of these methods to include informal online discourse as well as more edited text will increase the range of data sources that they can be applied to and may allow analysts access to the opinions of segments of society that are otherwise difficult to gauge.

Many of the challenges of the present task are analogous, though not 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. 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. These favorability judgments often interact in unexpected or counterintuitive ways. In the domain of U.S. politics, for example, it is likely that knowing a person's attitude toward abortion will help to inform a guess at that person's attitude toward the death penalty.

2.1 The `politics.com` discussion database

We created a database of political discourse downloaded from `www.politics.com`, a site devoted to discussion and debate on the topic of U.S. politics. The database consists of 77,854 posts organized into topic threads, chronologically ordered, and labeled according to the author and the author's self-described political affiliation. The posts are further divided 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 if it was identical to a substring in a previous post in the same thread by another poster. The database contains 229,482 individual text chunks, about 10 percent of which (22,391 chunks) are quotes from other posts.

There were a total of 408 unique posters identified in the dataset. The number of posts by each author follows a 'long-tailed' 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, with the second most prolific poster registering 3,801 posts.

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 6,481 politically oriented syndicated columns published online on right and left leaning websites `www.townhall.com` and `www.workingforchange.com` (4,496 articles and 1,985 articles, respectively). We also used a wordlist of email, chat and text message slang, including such terms as "lol," meaning "laugh out loud."

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, (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.

2.2 Political orientations

In this paper, we address the problem of political sentiment analysis as a kind of user classification: given a user's posting behavior, we want to assign them to a political orientation. First, then, we need to identify a range of possible political orientation labels which can be assigned to the users.

There is necessarily an element of arbitrariness in any selection of labels we might make. Our choice was motivated in large part by the kind of information we have available about the political orientation of individual users in our sample. In the `politics.com` dataset, political orientation labels are derived from users' self-descriptions given in their profiles. Users were allowed complete freedom in how they worded their self-descriptions, so some of the political affiliations needed to be normalized manually when the database was constructed. A description such as *true blue* was translated to *democrat*, whereas *USA Skins* was translated into *r-fringe*. Using these normalized self-descriptions, we arrived

at a classification including: *centrist, liberal, conservative, democrat, republican, green, libertarian, independent, l-fringe* and *r-fringe*.

For the experiments described in this paper, we took a simplified approach to political orientations. Users were grouped into three classes: LEFT (*liberal, democrat, l-fringe*), RIGHT (*conservative, republican, r-fringe*), and OTHER (*centrist, green, libertarian, independent*). A summary of the distribution is given in Table 1. Users listed as *unknown* either gave no political self-description, or gave a description that could not be normalized to one of the given classes.

take in Table 1

3 Sentiment analysis

To test the applicability of sentiment analysis methods to predicting user's political affiliation, we applied a variant of Turney's (2002) PMI-IR method. In Turney's original application, product reviews and other opinion-oriented texts are tagged for part of speech, and descriptive phrases are extracted using simple tag sequence templates. A large database of text (in Turney's case, the World Wide Web) is then searched to find occurrences of each descriptive phrase in the vicinity of the anchor terms *excellent* and *poor*. Based on these counts, Turney calculated the pointwise mutual information (PMI) of each descriptive phrase with each anchor term. PMI is an information theoretic measure for two events x,y of how far the true joint probability of the two events $P(x,y)$ differs from the joint probability that would be expected if the two events were independent:

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

If PMI(x,y)>0, then the events x,y occur more often that one would expect given their marginal probabilities $P(x)$ and $P(y)$, suggesting that the two events are associated. The "semantic orientation" (SO) of each descriptive phrase in a review is then the difference between the phrase's association with *excellent* and its association with *poor*:

SO(w)=PMI(w,excellent)-PMI(w,poor)

The overall orientation of a text is the average of the SOs of the phrases in the text. If the average SO of the descriptive phrases in a review is greater than zero, then the review is taken to be generally positive. When the average SOs were compared to the summary recommendations produced by the authors of the reviews, Turney's method yielded an 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 anchor terms *liberal* and *conservative*:

SO(w)=PMI(w,liberal)-PMI(w,conservative)

For this application, PMI scores for words were computed based on counts from the 200 million-word Reuters news corpus RCV1 (Lewis, Yang, Rose & Li. 2004). First 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. Some examples can be seen in Table 2: when viewing individual orientation values for phrases, the results were often (but not always) intuitive.

take in Table 2

We took a fairly straightforward approach to using this adaptation of Turney's method for user classification, by classifying texts for the 184 users who can be classed as either LEFT (Democrat, liberal, l-fringe) or RIGHT (Republican, conservative, r-fringe), by averaging the political SO of the descriptive phrases in their posts. When compared to the user's political self-descriptions, this yielded an accuracy of 40.76%. In comparison, simply assigning all users the most frequent label (LEFT) yields a baseline accuracy of 52.17%, and informal manual annotation experiments suggests a ceiling of 87.50%.

Clearly, the results of applying this Turney-inspired method in the present manner to political texts are less than encouraging. There are a few likely 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. In addition, RCV1 corpus consists of edited text collected in the late 1990's, and so is rather different in both style and content from the politics.com discussions. Since Turney's PMI-IR method is a form of unsupervised learning, a very large quantity of reference data is generally necessary to get good results, and no very large collections of informal political discourse are available to use as a reference

corpus. Also, the choice of the anchor terms *liberal* and *conservative* was necessarily somewhat arbitrary. We selected the most intuitive pair of terms we could think of to begin with. If results with this pair had been more promising, it would have been worth investigating the possibility of refining the results by incorporating different anchor terms. 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 (a point that we will return to in section 5)

4 Text classification

To test the effectiveness of standard text classification methods for predicting political affiliation, we used the Naive Bayes text classifier Rainbow (McCallum 1996) to predict the political affiliation of a user based on the user's posts. For each poster in the sample, a composite text was created by aggregating all of that user's posts, labeled with the user's political self-description. Using these composite texts as training data, prediction accuracy was determined by five-fold cross-validation. The NB text classifier gave an accuracy of 63.59% a modest (though statistically significant) improvement over the 52.17% baseline accuracy

There are a few possible explanations for the mediocre 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. 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 oriented 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 [2], augmented with the above-described list of political words and specialized computer-mediated communication vocabulary. 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 than the model without spelling correction.

A third possibility to account for 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 purely 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 indicate that a classifier based on textual features will perform better for frequent posters than for light posters. Unfortunately, simply collecting more posts will give us a larger database to train from but will not solve this 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.

5 Social network analysis

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. Each post is part of an on-going debate among the regular users of the site, and from users' posting behavior, we can locate their position within the political ecosystem of the site's participants.

Following similar reasoning, Efron (2004) used a combination of Turney-style sentiment detection and co-citation analysis to assign political orientation labels to a set of densely cross-linked weblogs. Unlike web pages, forum posts rarely contain links to other websites. However, many posts refer to other posts by quoting part of the earlier post and then offering a response, or by addressing

another user directly by name. We can use these explicit references to other users and their posts to create co-citation links between individual users.

In the `politics.com` dataset, the 41,605 posts by users classified as either LEFT or RIGHT contained 6,221 citations to other users who could also be classified as either LEFT or RIGHT. These citations consist either of using quoted material from an earlier post in the thread, or explicitly mentioning the name of another user (we excluded users whose names are homonymous with common English words to avoid false links). The majority of these links (77.5%) were to a user on the opposite side of the political spectrum. In this respect, citations between forum users appears to be markedly different from the inter-blog linking relationship discussed by Adamic & Glance (2005), who found that liberal and conservative blog sites largely link to other sites of the same political outlook. Thomas, Pang & Lee (2006) found similar patterns in Congressional floor debates, in which speakers tended to make explicit reference to other speakers who shared their general political orientation. Weblog communities, congressional debates, and online political forums are all highly polarized discourses, but the form that this polarization takes is quite different in the various genres. Perhaps the key difference is that blogs and formal political debates are intended for a mass audience of supporters, while informal political discussions are produced for the sole benefit of the participants themselves.

The conflict-oriented nature of the interaction also sheds some light on the failure of PMI-IR to reach even baseline accuracy. Participants in the discussion

raise issues that they see as damaging to the other side, leading to a pattern in which partisan figures and issues are more likely to be mentioned by posters on the opposite end of the political spectrum. For example, liberal-identified posters in our sample are highly critical of George W. Bush and the war in Iraq, while conservative-identified posters prefer to focus on Bill Clinton's shortcomings. This may be a partial explanation for the observation made in section 3 that the political orientation of the noun phrases in the texts is sometimes inversely related to the authors' political orientation.

To exploit this source of information, we used patterns of shared co-citations to group users into debating 'teams'. We first constructed a graph representing citation patterns, with each user represented by a node and each quoted post represented by an edge. The co-citation graph for one thread is given in Figure 3.

take in Figure 3

Given the simplified political orientation schema we are assuming, one might expect the connected components of the user co-citation graph to be bipartite graphs. However, since users sometimes cite people they agree with, and users may disagree on some issues with users who share their political orientation, the actual situation is much more complex. By hypothesis, though, users who play a similar rhetorical role in the discussions are likely to have similar opinions. Therefore, the more alike two users' citation profiles are, the more likely they are to share a political orientation.

To find groups of users with similar citation profiles, we first computed a low-rank approximation (via singular value decomposition) of the co-citation graph's adjacency matrix, to reduce noise and to highlight second-order structural generalizations (Drineas, Krishnamoorthy, Sofka & Yener 2004). We then computed the distance between each pair of users in the resulting 'citation space':

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

Based on these distances, we clustered the users to find groups of posters with similar citation patterns, using a single-link hierarchical clustering algorithm (de Hoon, Imoto, Nolan & Miyano 2004). Finally, the hierarchical cluster tree is cut at a depth determined by cross-validation to form groups of users with closely matched citation profiles. The clustering solution for one segment of the user database is given in Figure 4.

take in Figure 4

Note that the clustering hypothesis is not always correct: some users with very similar citation habits do not in fact share a self-assigned political orientation. However, for the most part, when users are assigned to a cluster, they are assigned to a cluster with common orientation.

To assign an orientation to these user clusters, we simply treat each cluster as if it were a single aggregate user. We gathered all the posts from all the users in

each cluster, applied the Naive Bayes classifier discussed above to the collected posts, and then assigned the predicted affiliation to all the users in the cluster. Since the combined posts of a cluster of users provides more evidence for the cluster's affiliation that the posts of any one user, we would expect the text classifier to perform better on clusters.

Indeed, this approach yielded more promising results than simply using Naive Bayes on individual users. For all users, this approach yields 68.48% accuracy, a significant improvement over straightforward Naive Bayes. And, for users with >500 words, this improves to 73.00% with clustering.

6 Future work

In this paper, we describe experiments using a number of well-known natural language processing techniques to predict the political orientation of posters involved in informal online discussions. A summary of the results is given in Table 3.

take in Table 3

A number of technical improvements could be made on the approaches we describe in this paper. Support vector machines could be used in place of Naive Bayes for text classification, and alternative clustering algorithms may be of help. Further information within posts may be available to improve the quality of citation detection. For example, a mention of a user's name might be a vocative expression which creates a citation link, or it might be simply a reference to that individual. Also, group identifiers, such as occur in the example:

It's very sad you conservatives can't win arguments on your own merit, you have to go petty and clone us liberals!

provide excellent evidence for a poster's orientation, and could be exploited more directly in text classification. Finally, the thread structure itself may also yield clues to better construct teams and find links.

In addition, 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. (Wilson, Wiebe & Hoffman 2005), Nasukawa (Nasukawa & Yi 2003), and Turney (Turney 2005). Variants of 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.

However, even without these extensions, the conclusion is clear: purely text-based methods (sentiment detection and text classification) performed relatively poorly at predicting user's political orientation. This is not surprising, given the nature of the texts. Since the posts are part of an on-going interaction, their meaning can only be understood within the context of a particular social environment and discourse situation. Unlike product reviews or news articles, they are not intended to be understood in isolation by a generic reader. Our results further suggest that information gained about the rhetorical relations between

posters and the roles they take in the discourse is of use in identifying the political sentiment of the posts.

In addition, what we are after here is a user's political orientation. A political orientation is not the same as a sentiment, nor is it a topic. It may be a bundle of sentiments with respect to a range of topics, but it is inherently more multi-faceted than what sentiment analysis or text classification algorithms have be designed to identify. 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.

Acknowledgments

This work is supported by KAKENHI 18700158, Grant-in-Aid for Young Scientists (B) provided by the Minisitry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

Adamic, L. & Glance, N. (2005) "The political blogosphere and the 2004 U.S. election: Divided they blog." *Proceedings of the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. Chiba, Japan.

de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. (2004) "Open source clustering software." *Bioinformatics* **20**, 1453–1454.

Drineas, P., Krishnamoorthy, M., Sofka, M. & Yener, B. (2004) "Studying e-mail graphs for intelligence monitoring and analysis in the absence of semantic information." 'Intelligence and Security Informatics', Vol. 3073 of *Lecture Notes in Ccomputer Science*, Springer-Verlag, pp. 297–306.

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

Efron, M., Zhang, J. & Marchionini, G. (2003) "Implications of the recursive representation problem for automatic conceptidentification in on-line governmental information." *Proceedings of the ASIST SIG-CR Workshop*.

Lewis, D., Yang, Y., Rose, T. & Li., F. (2004) "RCV1: A new benchmark collection for text categorization research." *Journal of Machine Learning Research* **5**, 361–397.

Malouf, R. & Mullen, T. (2007) "Graph-based user classification for informal online political discourse.", *Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW).*

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. & Fukushima, T. (2002) "Mining product reputations on the web." Proceedings of the eighth ACM

SIGKDD international conference on Knowledge *discovery and data mining*, pp. 341 – 349.

Mullen, T. & Collier, N. (2004) "Sentiment analysis using support vector machines with diverse information sources." *Proceedings of EMNLP*.

Mullen, T. & Malouf, R. (2006) "A preliminary investigation into sentiment analysis of informal political discourse." *Proceedings of the AAAI-2006 Spring Symposium on "Computational Approaches to Analyzing Weblogs"*.

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

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

Thomas, M., Pang, B. & Lee, L. (2006) "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts." *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Turney, P. (2002) "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th*

Annual Meeting of the Association for Computational Linguistics, ACL, Philadelphia, Pennsylvania, pp. 417–424.

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

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

Wilson, T., Wiebe, J. & Hoffman, P. (2005) "Recognizing contextual polarity in phrase-level sentiment analysis." *Proceedings of HLT-EMNLP*.

[1] http://www.abqtrib.com/albq/op_editorials/article/0,2565,ALBQ_19867_4604117,00.html

[2] http://aspell.net/

EDITORIAL: Time is now to raise city's minimum wage

The Albuquerque Tribune

Friday, April 7, 2006

Here we go again, and hopefully three's the charm, because the hard-working people of Albuquerque have earned a little help. More important, they need it—some desperately. After recent failures in Albuquerque and statewide to establish a minimum wage above the failing federal minimum of $5.15 per hour, a revitalized city effort looks promising: Albuquerque Mayor Martin Chavez has announced his intention to establish a citywide minimum wage within the next few months. City Council President Martin Heinrich has proposed an ordinance that appears to have the necessary votes for passage and that would gradually raise the minimum wage each January—starting with $6.75 next year, until it reaches $7.50 an hour in 2009.

Figure 1: An opinion expressed in traditional editorial style

- LOL. If inflation won't happen with minimum wage increases, then why not put minimum wage to a 100$ an hour? Don't worry it won't affect prices....the extra money for the wages will come out of the sky.

- What's the matter rockhead? You don't want to help the poor needy people? You rich bastard koolaid drinker!

- that was pretty stupid man honestly...grow up !

- I don't have a kid brother

Figure 2: Opinions expressed in informal language, from the politics.com data set

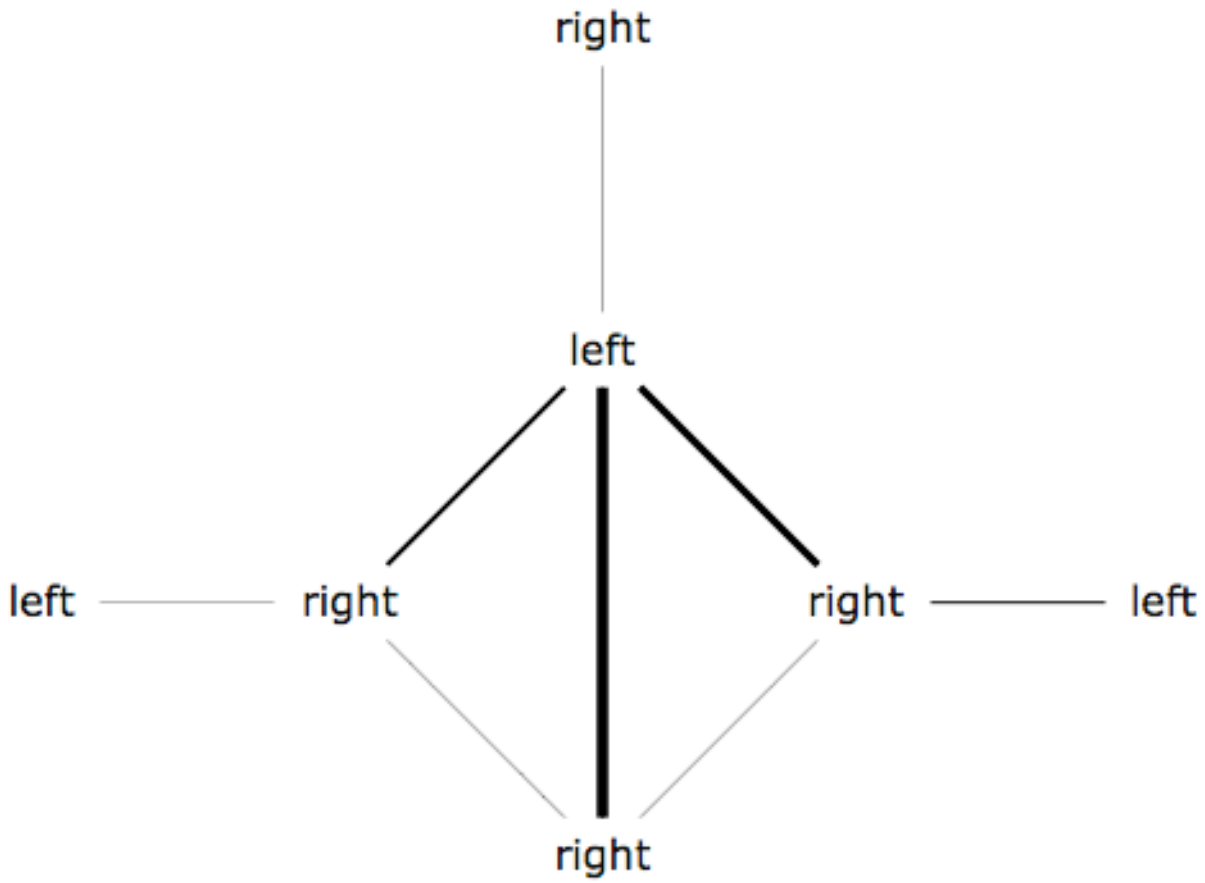

Figure 3: Co-citation graph for one thread in `politics.com` dataset. Nodes represent individual posters (labeled with the user's political affiliation), and darker edges indicate greater citation frequency.

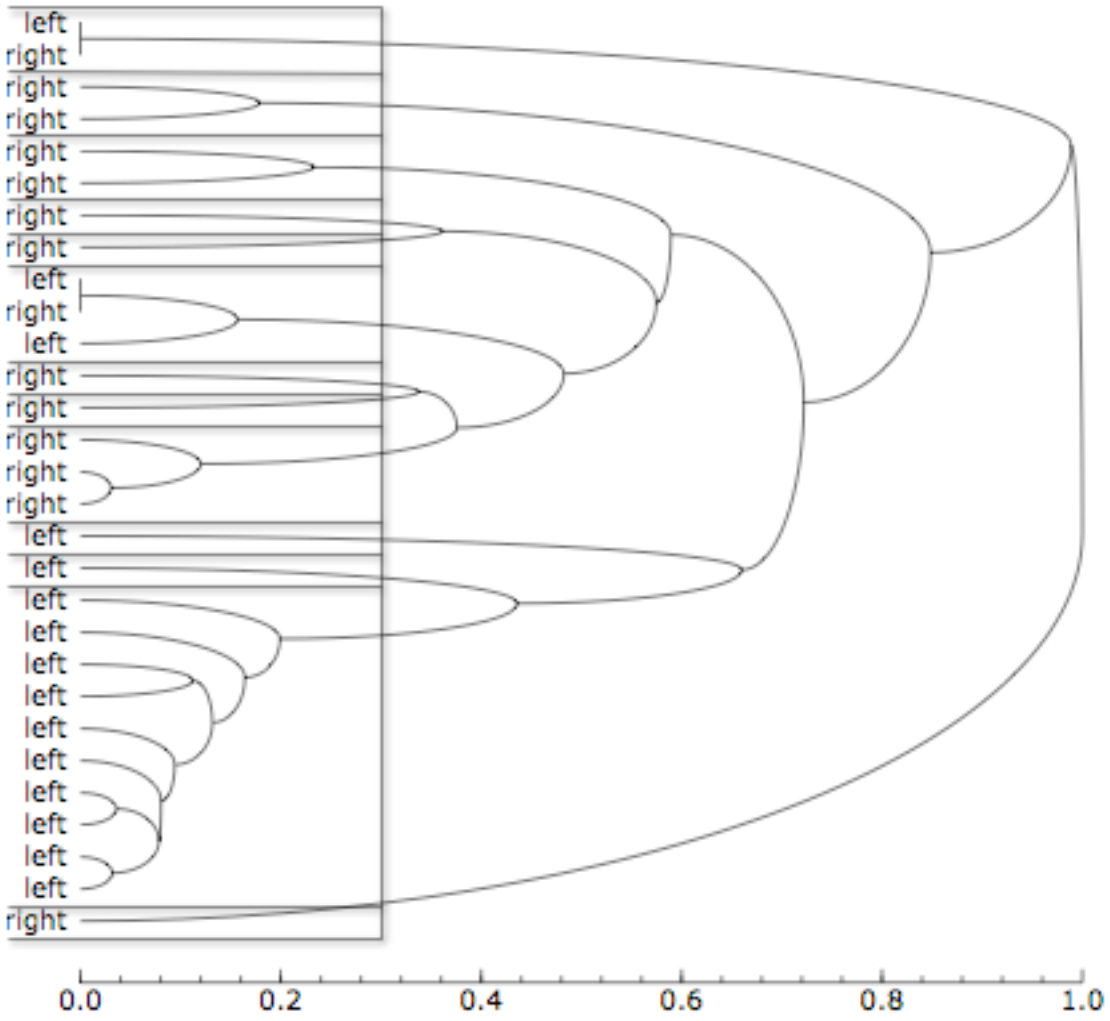

Figure 4: Hierarchical clustering based on co-citation patterns. Users are labeled with their true self-reported political orientation.

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

Table 1: Distribution of users in the `politics.com` dataset by general class and by normalized self-descriptions.

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		

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

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 reader	87.50	91.00

Table 3: Summary of results