BLEWS: Using Blogs to Provide Context for News Articles

Michael Gamon*, Sumit Basu*, Dmitriy Belenko*,
Danyel Fisher*, Matthew Hurst+, Arnd Christian König*
*Microsoft Research and +Microsoft Live Labs
One Microsoft Way
Redmond, WA 98052
+01 1-800-642-7676
{mgamon; sumitb; dmitryb; danyelf; mhurst; chrisko}@microsoft.com

Abstract

An overwhelming number of news articles are available every day via the internet. Unfortunately, it is impossible for us to peruse more than a handful; furthermore it is difficult to ascertain an article's social context, i.e., is it popular, what sorts of people are reading it, etc. In this paper, we develop a system to address this problem in the restricted domain of political news by harnessing implicit and explicit contextual information from the blogosphere. Specifically, we track thousands of blogs and the news articles they cite, collapsing news articles that have highly overlapping content. We then tag each article with the number of blogs citing it, the political orientation of those blogs, and the level of emotional charge expressed in the blog posts that link to the news article. We summarize and present the results to the user via a novel visualization which displays this contextual information; the user can then find the most popular articles, the articles most cited by liberals, the articles most emotionally discussed in the political blogosphere, etc.

Keywords

Blogosphere, polarity and emotion detection, news articles, emotion detection, visualization.

Introduction

In recent years, a huge variety of news sources have become available to us via the internet. Whereas in the past we were limited to only local sources and a few national/international papers, we are suddenly able to read articles from thousands of news organizations. As it becomes impossible to keep ahead of such a cornucopia of news, aggregator sites like Google News make things more manageable by clustering articles into top stories. However, this presents us with a new problem—how are we to know which articles to read? For a given topic, there may be thousands of articles from hundreds of sources, but at best we only have time to read a few. Furthermore, if we do read a particular article, we read it through a pinhole, having no idea whether it is important or representative. Having broken free of the editorial bias of a particular newspaper, we are now overwhelmed by the number of choices for each topic and the lack of context for each article.

Fortunately, at the same time, a potent source of editorial content has been growing faster than the information sources themselves—the blogosphere. Of course, were we to only read one or more blogs to get our news, we would again be trapped in their particular editorial biases. However, there are other ways in which we can leverage their editorial efforts, since there are many ways in which bloggers give editorial feedback on news articles. First, the mere act of linking to stories is an important cue as to the relative value of the story. Second, the political orientation of the blogger's behavior can tell us a lot. Finally, the way in which the blogger talks about the story in their blog post can give us valuable additional information: is this article contentious or is it received more neutrally in the blogosphere?

In this paper, we seek to harness this contextual information to give users context for each story they consider reading: how popular is the story, what kinds of people are linking to it, and how they feel about it. These are all abstract notions that could be quantified in many ways: in this work, we choose a metric for each category that is both informative and can be reliably computed or inferred. Furthermore, we focus exclusively on the domain of political news. For measuring popularity, we make use of the number of blog posts from political blogs (gathered from public lists) linking to the article, after first collapsing articles with greatly overlapping content. For determining the political orientation of the bloggers, we use readily available public lists. Finally, as a means for assessing how the blogger feels about the article, we develop a classifier with which we can estimate the level of emotional charge of the post.

Given this information, we can ask a variety of questions of the news, from the relatively mundane "what is the most popular article" to "which articles about health care are equally popular between conservatives and liberals" or "which articles are liberals most emotional about?" In addition to such queries, though, we wish to present the user with this contextual information in a digestible form, and as such we develop a user interface to summarize and visualize this information along with news results. We will show how this information can also be used for sorting articles along dimensions of interest to address some of these questions.

Note that our approach does not depend on fine-grained analysis of any *particular* post—we instead report on the *aggregate* statistics of the data. We also do not expose the

particulars of our classification results on any particular blog post at the top level of the interface. This is an important point, since it allows the interface to be reasonably robust to errors in our classifiers our in human labels. For instance, if some blog is mistakenly labeled as conservative instead of liberal, or we misestimate the level of political charge in a post, it will have little effect if this is one blog and one post amongst the fifty blog posts pointing to the article. As long as our classifiers and labels are reasonably accurate, the sheer mass of evidence will outweigh the noise to give the user a good sense of the actual content.

We have prototyped our approach in a system called BLEWS (**Bl**ogs and N**ews**). In the remainder of this paper, we detail the methods we have developed to approach these goals. We begin by discussing other work that has addressed the connections between blogs and news. We then describe our data collection infrastructure, followed by our algorithm for collapsing news articles with mostly overlapping content. Subsequently, we present statistics on the linkage between political blog posts and news articles in a given time window. Next we describe how we estimate the various contextual quantities for each article, along with a discussion of why emotional charge has the appropriate balance of informativeness and reliability. We then describe how we provide summarization and visualization of these features. Finally, we conclude with a discussion and thoughts on future work.

Related Work

Various research has been reported addressing certain aspects of context with respect to news and social media. Perhaps the most immediate is TextMap (Godbole, Srinivasaiah, & Skiena 2007) which analyzes news and blogs for a number of features such as sentiment around extracted entities (e.g. politicians) and tracks these features over time. Work on the political orientation of blogs, and the relationships and dynamics between political blogs is reported in (Adamic & Glance 2005). This work, however, doesn't address the relationship—an ecological relationship—with the news media. Viewing the relationship between social media and news media as ecological is an important distinction from much research in social media to date which characterizes the blogosphere itself as ecological. Pertinent to our study of the relationship between social media and mainstream news is (Zuckerman 2007) which studies the relative biases of different news sources geographically. A number of demonstration systems exist which visualize the news in novel ways, providing additional context helping the user better understand the events. A good example is (Buzztracker.com 2007) which provides geographic context to news stories based on the location collocated in articles. We will discuss related work in sentiment detection in the context of the corresponding section of this paper.

Data Acquisition

The work described in this paper takes advantage of a social media analysis platform we are developing. The platform is designed to provide uniform access to a real time stream of a variety of social media content including weblogs, usenet, microblogs (such as Twitter, Jaiku and Pownce) and message board forums. In addition, the platform offers a content store, and mining infrastructure to support both real time and offline mining.

The weblog acquisition component adopts standard approaches to weblog crawling: it monitors ping servers and crawls feeds in response to ping events. For blogs that do not provide regular pings, it performs scheduled crawling. Partial feeds are augmented with an intelligent scraping mechanism which parses the structure of the permalink page, extracting the complete content of the post.

Weblog posts are made available as a stream and via interaction with the content store. For the BLEWS system, we acquire the stream of blog posts and filter for posts that contain news links, storing them in a database. Offline processing then identifies posts from political blogs and detects emotional charge around the news links in the post, storing that information in the database.

We maintain a whitelist of more than a thousand known labeled political blogs, aggregated from Wonkosphere.com and our own crawl and classification based on link structure with respect to known blogs with known orientation. All blog entries are from blogs on this whitelist. This allows us to maintain reasonable certainty about blog political orientation.

Detecting Duplicate Articles

Many news articles are based on syndicated stories from the Associated Press or Reuters agencies. Articles that are based on the same syndicated content very often show large or even complete overlap in content. The benefit of our approach comes from being able to aggregate information about in-links, emotional charge and attributes of the in-linking blogs for individual articles. Consequently, we need to be able to identify multiple occurrences of the same article, which is non-trivial for a number of reasons.

In many cases, non-identical blog-links will point to the same article on a particular news source—for example, links to many news sources are insensitive to parts of the domain label being omitted (e.g., '*www.nytimes.com/X*' and '*nytimes.com/X*' generally lead to the same content); similarly, many articles can be referenced either via their location in the content hierarchy (e.g. '*www.newssource.com-/news/international/...*') or a content identifier (e.g. '*www-.newssource.com/news.aspx?contentguid=X145739*'). On the flip side, links with identical URLs will not necessarily refer to the same article—for example, a link to the "latest" editorial by a columnist will not lead to the same article a month later. Hence, we need to consider the actual text of an article pointed to by a new blog post when resolving which links refer to identical articles.

Typically only a subset of the text on each crawled web page corresponds to the actual news content, whereas the remainder may contain information like copyright statements, advertisements, links to other articles within the same news source, etc. While our news acquisition module contains a news extraction module to eliminate this content, we have to account for the extraction process being less than perfect. Moreover, some of this content is generated automatically,

so it may vary even for consecutive accesses to identical URLs. Similar issues occur for duplicate articles posted on different news sources which may have different news templates and copyright notices, such as an Associated Press article syndicated across newspapers. Ideally, we would like to disregard all of this "meta-content" for the purpose of identifying duplicate articles.

The issue of detecting duplicate documents in large text corpora has received significant attention (e.g. (Fetterly, Manasse, & Najork 2005)(Huffman *et al.* 2007)(Schleimer, Wilkerson, & Aiken 2003)) recently. The focus of these approaches has been reducing the computational and space overhead required to detect duplicates through the use of fingerprinting/shingeling techniques. Our approach is orthogonal to these in that we are concerned with identifying the right substrings within news pages to use for duplicate detection, separating the actual news article from other text content. The techniques mentioned above can be combined with our approach to efficiently filter out documents which cannot be duplicates of each other as an initial processing step.

The insight that we leverage to distinguish news content from other content is as follows: whereas differentiating the two within a single article is often difficult, we can leverage the property that much non-news content (article links, copyright information, columnist bios) is common across different news articles from the same news source, whereas the article text itself is not. We leverage this property by identifying non-news context by its frequency in the news corpus. To do so, we compute the document frequency of any (maximal) n-gram (for variable values of n) within the news corpus. As short n-grams may be common independently of whether if they correspond to non-news content or not, we require a minimum length n ($n \geq 10$ in our prototype). The n-grams we discover in this way now allow us to identify parts of articles correspond to common text (e.g. copyright notices) and which parts are rare, allowing us to focus on the rare text documents have in common for duplicate detection.

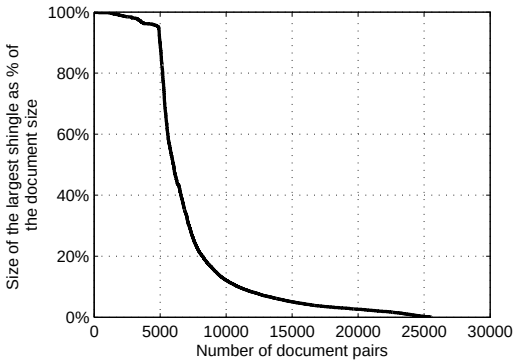

Figure 1: Distribution of maximum n-gram size.

To compute n-gram frequencies within the news corpus, we use suffix arrays. A suffix-array is a compact representation of a suffix tree (McCreight 1976) over a corpus, with the key advantage that its space-consumption does not grow with the size of the underlying vocabulary. Once a suffix-array over the corpus T of news articles containing C words has been constructed, we can compute the document frequency of all maximal n-grams in $O(C \log C)$ time using the algorithm of (Yamamoto & Church 2001), while requiring $O(C)$ space, which in practice means approximately $6C$ bytes for the suffix array (4 bytes per word) and the lcp array (2 bytes per word). Combined with the fact that we can restrict the duplicate detection to a subsets of documents (e.g. we only look for duplicates in news articles that were referenced in blog entries we ingested less than 6 weeks apart) this means we can store the text and suffix array in main memory, making the duplicate detection itself very fast.

As we have described before, we are only interested in n-grams that are long (i.e., in our prototype $n \geq 10$); we divide these into two categories: *rare n-grams* that occur in few documents (e.g., less than 50) and *frequent n-grams*, which are more common. Now, by iterating over all such n-grams, we compute the following for all articles: (a) $S(a_i, a_j)$ = the size (in number of words) of the largest rare n-gram articles a_i and a_j have in common, and (b) $W(a_i)$ = the substring of all words in an article a_i not covered by frequent n-grams. We now tag any pair of documents a_i, a_j for which the following conditions hold as duplicates; let a_i denote the larger of the two documents and $size(a_i)$ its size (in words):

- The ratio $\frac{S(a_i,a_j)}{size(a_i)}$ exceeds a threshold α.
- The fraction of words in $W(a_i)$ that are covered by at least one rare n-gram occurring in both documents exceeds a threshold β.

The precise value of these thresholds may vary depending on the nature of the news sources used; however, there is some indication that our method is not very sensitive to small changes in the thresholds. To illustrate this, we have plotted the distribution of the values of $\frac{S(a_i,a_j)}{size(a_i)}$ for all document-pairs from a set of 40K news articles below in Figure 1: the 5K document-pairs with the largest values of this ratio overlap almost completely, but subsequently the degree of overlap drops off fast, meaning that changing the threshold α on document-overlap from 95% to 80% will result in little change in the output.

Links between Blog Posts and News Articles

The data discussed in this section is based on blog post acquisition from November 20-29. The link structure between blog posts and news articles is dense enough to demonstrate the benefits of our approach. In this timeframe we acquired a total of 4494 blog posts from 473 distinct political blogs. These posts linked to 4923 unique news URLs. After duplicate detection, this number dropped to 4665. In order to analyze the link structure between blogs and news articles, we represent the information as a directed graph: blogs and news articles are nodes in that graph, each link from a blog to a news article is a directed edge. The distribution of in-links to news articles and outlinks from blog posts is shown in Figure 2 and Figure 3.

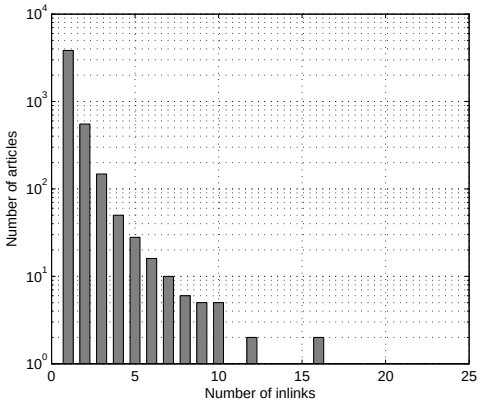

Figure 2: Distribution of number of inlinks on news articles (Nov 20-29)

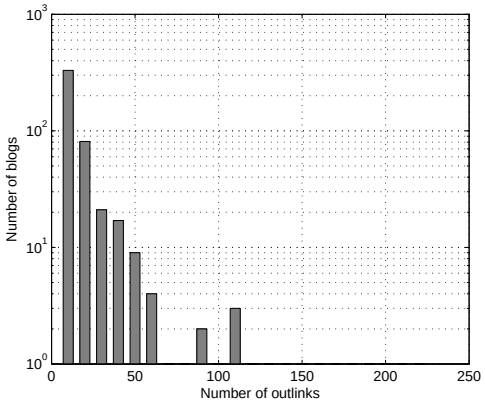

Figure 3: Distribution of number of outlinks to news articles per political blog (Nov 20-29).

The most active blogs in our acquisition are listed in Table 1, the most linked-to news articles in Table 2. A visualization of the graph is shown in Figure 4, where blogs are represented and located according to their orientation as blue (liberal), red (conservative) and green (independent) nodes. News articles are white nodes. In this graph, we note that while there are some contentious articles in the middle, there are also a great many articles that seem to be interesting only to the left or the right.

HuffingtonPost	211
The BRAD BLOG	150
Norwegianity	121
Wonkette	113
Think Progress	107
News About Ron Paul	106
First Read	104
NewsBusters	85
Huckabee 08 Unofficial Blog	81
Instapundit (v.2)	72
Prometheus 6	68

Table 1: Blogs ranked by number of posts that contain links to news articles (Nov 20-29)

Mansoor Ijaz: A Muslim Belongs in the Cabinet (Christian Science Monitor)	23
Dan Balz and Jon Cohen: Huckabee Gaining Ground in Iowa (Washington Post)	13
Donald Lambro: Study: Democrats the Party of the Rich (Washington Times)	13
Sara A, Carter: Terrorists target Army base—in Arizona (Washington Times)	12
Marty Griffin: Wounded Soldier: Military Wants Part of Bonus Back (KDKA—Pittsburgh)	12
Trent Lott announces his resignation (NBC)	11
Robert Novak: The False Conservative (Washington Post)	11
Andy Soltis: Blame U.S. for 9/11 Idiots in Majority (New York Post)	11

Table 2: News articles (Nov 20-29) with the most inlinks from political blog posts

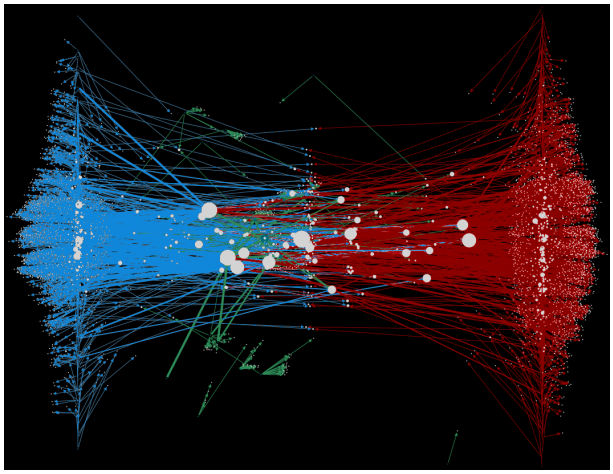

Figure 4: Graph of links from conservative, liberal and independent blogs to news articles.

Detecting Emotional Charge

Most of the work in sentiment detection and classification has been centered on the distinction between positive and negative sentiment (for example (Pang, Lee, & Vaithyanathan 2002), (Turney 2002), (Kim & Hovy 2004)). This distinction offers distinct practical advantages: in many scenarios (such as product reviews, for example) it is both a useful and relatively clear cut classification task. There has also been a body of work concerned with the distinction between subjective and objective text (for example (Wiebe *et al.* 2004)), which has applications for example in information retrieval: a user interested in factual information will benefit if opinionated and subjective texts and passages are avoided as targets for extraction. Mishne uses user-annotations on Livejournal posts to classify these posts into 132 categories of mood, and also visualizes the user annotations ((Mishne 2005), (Mishne *et al.* 2007)). The categories are designed to capture personal states in detail, examples are "accomplished", "thirsty", "worried". Balog and de Rijke (Balog & de Rijke 2006) analyze the temporal distributions of these mood annotations. We believe that neither of these strategies is very suitable for the analysis of political blog post text referring to news links. Political blog posts are in the vast majority expressing opinion or offering political commentary; as such, both the subjective versus objective distinction is mostly irrelevant, as is the LiveJournal notion of mood. A complete and sound treatment of sentiment in political discourse would have to embrace at minimum the following distinctions:

a. Sentiment: the agent of the evaluative attitude, the object of the evaluative attitude and the polarity and strength of the attitude.

b. Argumentative structure: is an article cited in support of the author's argument, or in opposition to it?

(b) has been tackled in scientific writing (Teufel 2006), to the best of our knowledge it has not received any attention in less structured writing such as blog posts. In blog posts, it is a very difficult task: in a pilot study, we attempted to hand-label agreement of blog entries with the article they pointed to, and found that it was almost impossible for the annotators to reach consensus.

To illustrate the challenges in automatically analyzing blog posts according to sentiment and argumentative structure, we have collected some examples below.

- *Taking a break from not getting anywhere on Iraq, Congress looks to not get anywhere on domestic issues for a little while. [news link]*—Negative sentiment towards a state of affair, news article cited in support.
- *If you liked last term's Supreme Court, you're going to love the sequel. [news link]*—Negative sentiment towards a state of affair expressed in ironic disguise.
- *Honestly: you can't make this stuff up. [news link] Michael Ledeen at The Corner: "The Left hates Rush, above all-as in the case of Thomas-because of the quality of his mind and the effectiveness of his work." The quality of his mind? The quality of his mind????!! I mean, we are talking about Rush Limbaugh, right?*—Negative sentiment towards an opinion stated in a news/commentary link and towards a public figure.
- *Leftard policy at its finest. $100,000 a year and they're in public housing? [news link] I am shocked the Washington Post would even report it.* Negative sentiment expressed towards a state of affairs as reported in the news link, negative sentiment in ironic disguise towards the news provider.
- *MoveOn.org betrayed us, or more specifically, they betrayed the Democrats, many of whom, including Hillary, betrayed us, the American military and General Petraeus. Just thought I'd point that out [news link].*—Negative sentiment towards political organizations and persons, news link cited as factual support.
- *More evidence of a turning of the tide [news link], from an editorial in the Washington Post: (long quote from news article). Obviously, I don't expect Hillary-or any other Democrat (except Lieberman, of course)-to take note of these positive developments. In the words of Bush Senior, "couldn't do that, wouldn't be ~~right~~ good politics."*— Positive sentiment towards a state of affairs as reported in the news article, negative sentiment towards political figures.
- *The Wall Street Journal [news link] reports that the Republican Party is losing its traditional support from the business community. Factors cited include opposition to the war, liberal attitudes on social issues, and opposition to the fiscal responsibility of the Republican Party. Many also disagree with the Republican denial of global warming as "some business people want more government action on global warming, arguing that a bolder plan is not only inevitable, but could spur new industries."*—Neutral summarization of a news article.

Given these challenges, we decided to target a different quantity, which we call "emotional charge." This is similar in spirit to the work in (Wilson, Wiebe, & Hwa 2004), but in contrast to the research reported there we cast the problem as a binary decision between "emotionally charged" and "not emotionally charged" as opposed to a finer-grained ranking. Emotional charge can roughly be defined as the degree of excitement and agitation of the author at the time they are writing the text. We regard the text as the observable manifestation of this emotional charge. Of course, this notion is not without its problems either—for example, the author could be emotionally neutral at the actual time of writing, but pretending that they are enraged. Overall we believe, however, that there is a reasonable chance to infer from the text an indication of the emotional state of the author. Emotional charge thus defined cuts across both the positive/negative/neutral distinction and the subjective/objective distinction: both positive and negative emotion can be expressed in a forceful, emotionally charged way and in a matter-of-fact way. Similarly, a text can be subjective, but written by a level-headed author in a cool and detached manner. Finally, emotional charge is agnostic to the difficult task of identifying argumentative structure: Whether an article

is cited in support of or in opposition to the author's view, emotional charge is only concerned with the emotional state of the author. The notion of emotional charge is a simplification, but there are arguments for its soundness and usefulness. While the literature is rife with differing definitions of emotion (see e.g. the summary in (Kleinginna & Kleinginna 1981)), there is no doubt that emotion and its expression has a dimension of strength to it. Our notion of emotional charge, for example, corresponds to Scherer's (Scherer 1984) axis of "activity" in his typology of affective states. We also strongly believe that emotional charge is a dimension that is of interest in political discourse: heated discussion of a news topic in the blogs from a particular political camp is useful background information for a user choosing between news stories.

For our system, we construed the problem of detecting emotional charge from text as a supervised learning task. For an initial exploration of the task, an independent annotator manually categorized 600 randomly selected political blog posts containing embedded news links. 415 of these posts were categorized as emotionally charged, 131 as not emotionally charged, and 54 posts were eliminated either because they contained spam or had no political content. We trained a maximum entropy classifier on unigram feature vectors representing the blog posts. On tenfold cross-validation, the classifier achieved accuracy of 80.92%, over a baseline of 75.74% (assigning the majority class label: emotionally charged). This result is statistically significant at the 0.999 level, as determined by the McNemar test. The F-measure is 88.71 over the baseline of 86.19. Precision and recall for the detection of emotional charge is shown in Figure 5.

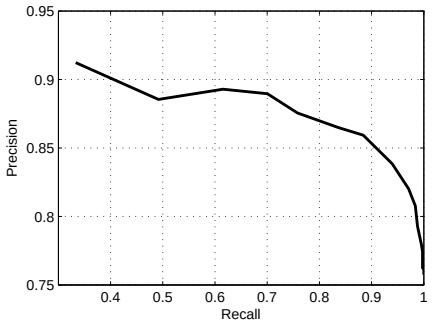

Figure 5: Applying the emotional charge classifier to blog posts.

Since the data set of 546 annotated blog posts (after removal of spam and irrelevant posts) is extremely small for a text classification task, we also investigated how much we could potentially gain from additional annotated data. For this purpose we conducted learning curve experiments, using decreasing subsets of the training data to train our classifier. Results of these experiments on tenfold cross-validation are shown in Figure 6.

As is evident from the angle of the learning curve, we can be reasonably confident that the addition of more annotated data will substantially boost our accuracy.

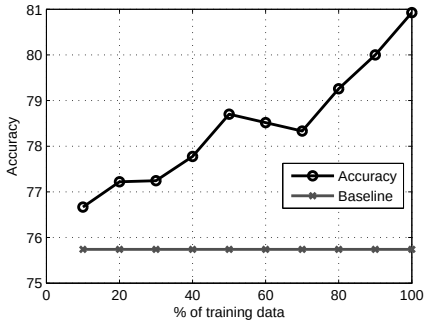

Figure 6: Learning curve on annotated data.

In a separate experiment, we tried to leverage available data from outside the political domain. As exemplars of generally non-charged text we used a combination of English Encarta encyclopedia text and the Reuters newswire. For emotionally charged text, we utilized a large set of web car reviews (450k reviews), customer feedback text and Pang and Lee's (Pang, Lee, & Vaithyanathan 2002) movie review data set. In order to overcome the problem of different domain vocabulary, we eliminated all features that were highly predictive of individual domains, using Dunning's log likelihood ratio (Dunning 1993) as a measure of predictiveness. Once the feature set was reduced in that manner, we trained a maximum entropy classifier to distinguish the charged from the uncharged text. The resulting "generic" classifier was then tested on the annotated blog posts from the political domain. Precision and recall are plotted in Figure 7. We

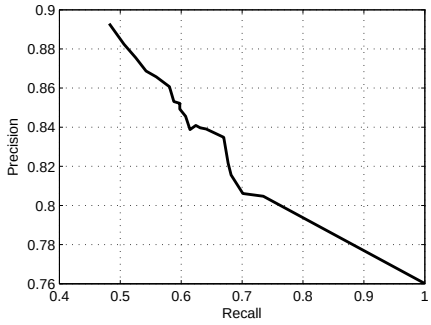

Figure 7: Applying the "generic" emotion classifier to blog posts.

believe that these initial experiments demonstrate that detecting emotional charge is a tractable problem. The next steps in our experiments will be:

- increase the amount of manually annotated blog posts
- train weights for a weighted combination of the generic and blog post classifier
- utilize additional manual annotation that indicates for each pair of blog posts the ranking in terms of emotional charge to train a ranker (e.g. RankNet (Burges *et al.* 2005))

Visualization and User Experience

The last component of the BLEWS system is the user interface, which is intended to convey the social context around the article, as discussed above. As a result of the emotional annotations described in the previous section, the interface uses as its raw data the annotated list of all in-links to all articles; each in-link is characterized by orientation (liberal or conservative), emotion, and a certainty level. These many values could be conveyed in many ways. We wanted to ensure that the information was conveyed in a way that was easy to read at a glance, even without particular training. We began with the following questions: users should be able to tell whether more liberals or conservatives had pointed to an article, and whether the reactions on each side were emotional or not. We wanted users to be able to easily find the most liberal (or conservative) article, and to be able to find the articles that generate the highest emotion.

We wanted to ensure that the interface would primarily reflect story *popularity*—that is, the total number of in-links—but that it could be easily used to get polarity (i.e., conservative vs. liberal), and that it could also be annotated by the level of emotional charge. The emotion value cannot be portrayed with unwarranted precision, as that would suggest more precision than we are able to estimate. Our solution is the "glowing bars" visualization suggested in *Figure 8*. The article headline is shown, surrounded by two indicator boxes; one indicator shows the number of liberal references, the other shows conservative references. Articles are ordered from top-to-bottom by their popularity. The indicator boxes are sized relative to the number of in-links: an article that has many liberal in-links has a large left-side blue wing, while an article with many conservative links has a large right-side red wing. Because the boxes contain precise article counts, users can easily evaluate articles either by relative or absolute weight.

Emotion is more difficult to convey. We aggregate emotion by simply taking the percentage of in-links that our classifier claims are better than chance—that is, over 50%—likely to be emotionally weighted. We convey this percentage as a 'glow' around the number: an article with many emotionally-loaded references has a large halo, while unweighted references have little or no glow. This allows users to easily compare glows in a relative sense without suggesting excessive precision. In Figure , for example, the articles are sorted by the number of liberal in-links. Note that the fourth article, "Trent Lott Announces His Resignation", has an emotional response from liberal in-links, but not conservative ones. A tooltip allows the user to explore the in-links to an article in more detail. In Figure , the user is examining the conservative "6" for the article "A Muslim belongs in the cabinet"; the five articles in the tooltip are the conservative in-links. The glow around some of the articles indicates in-links that our system labeled as emotionally-charged.

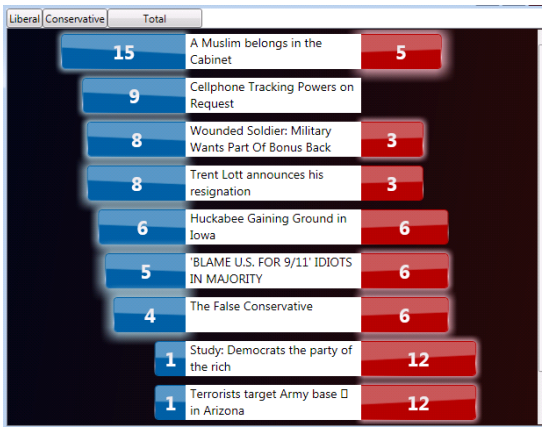

Figure 8: The BLEWS User Interface. "Wings" on each side of the title show the number of in-links to each article; the glow around each article suggests its emotional sentiment. In this view, the display is sorted by number of liberal in-links.

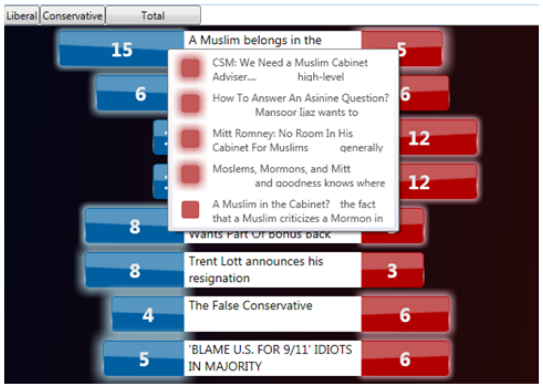

Figure 9: The Blews User Interface, showing in-link details for the first article in the display. Note the fuzzing around the blog entries, which suggest emotional affect.

Future work

The BLEWS system is a first foray into using a combination of text processing and link analysis to provide contextual information for news articles. In this prototype we have limited ourselves to a detection of emotional charge around links to news articles in posts from political weblogs. Political weblogs in our system are identified based on a whitelist of blogs with known political orientation. The information in the acquired data is much richer, however, and in future versions we plan to exploit this information in a number of ways. Link structure between blogs, for example, can be utilized to dynamically update the set of known blogs and infer political orientation. Similarly, political commentary around news stories is not limited to the active and established political blogs. It can also be found in other sources such as usenet, microblogs and personal blogs. In order to make use of this context, one would first have to separate posts with political content from posts with personal or other content. Text classification techniques can be used for this task. Political orientation could be classified from a combination of link information and textual content.

Another dimension that we have disregarded for our first prototype are the notions of "authority" and "influence". Once the link structure from blog posts to news stories be-

comes denser through incorporating more and more political blogs and individual posts from personal blogs, it becomes important to separate the lone ranters from those bloggers who are read and linked to extensively (as, for example, described in (Fiore, Tiernan, & Smith 2002)).

In addition, clustering techniques can be used to group news stories and blog posts according to topics, adding additional structure to the context around the news stories, and allowing the user to navigate topics spaces along with the contextual information on the news stories. Finally, we would like to explore the possibilities for creating a more complex model of emotional charge, using additional features and sources of information outside of the text itself.

Acknowledgements

We would like to thank Dave Steckler, Alexey Maykov, and Andre Mohr for their invaluable contributions to the data acquisition system and social media platform.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, 36–43. New York, NY, USA: ACM.

Balog, K., and de Rijke, M. 2006. Decomposing bloggers' moods—towards a time series analysis of moods in the blogosphere.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, 89–96. New York, NY, USA: ACM.

Buzztracker.com. 2007. Buzztracker: World news, mapped.

Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.

Fetterly, D.; Manasse, M.; and Najork, M. 2005. Detecting phrase-level duplication on the world wide web. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 170–177. New York, NY, USA: ACM.

Fiore, A. T.; Tiernan, S. L.; and Smith, M. A. 2002. Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, 323–330. New York, NY, USA: ACM.

Godbole, N.; Srinivasaiah, M.; and Skiena, S. 2007. Large-scale sentiment analysis for news and blogs.

Huffman, S.; Lehman, A.; Stolboushkin, A.; Wong-Toi, H.; Yang, F.; and Roehrig, H. 2007. Multiple-signal duplicate detection for search evaluation. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 223–230. New York, NY, USA: ACM.

Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, 1367. Morristown, NJ, USA: Association for Computational Linguistics.

Kleinginna, P. R. J., and Kleinginna, A. M. 1981. A categorized list of emotion definitions, with suggestions for a consensual definition.

McCreight, E. M. 1976. A space-economical suffix tree construction algorithm. *J. ACM* 23(2):262–272.

Mishne, G.; Balog, K.; de Rijke, M.; and Ernsting, B. 2007. Moodviews: Tracking and searching mood-annotated blog posts.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005*. SIGIR, ACM.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Scherer, K. R. 1984. Emotion as a multicomponent process: A model and some cross-cultural data.

Schleimer, S.; Wilkerson, D. S.; and Aiken, A. 2003. Winnowing: local algorithms for document fingerprinting. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 76–85. New York, NY, USA: ACM.

Teufel, S. 2006. *Argumentative Zoning for Improved Citation Indexing*. Springer.

Turney, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, 417–424. Philadelphia, US: Association for Computational Linguistics.

Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; and Martin, M. 2004. Learning subjective language. *Comput. Linguist.* 30(3):277–308.

Wilson, T.; Wiebe, J.; and Hwa, R. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, 761–769. San Jose, US: AAAI Press / The MIT Press.

Yamamoto, M., and Church, K. W. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Comput. Linguist.* 27(1):1–30.

Zuckerman, E. 2007. Global attention profiles - a working paper: First steps towards a quantitiative approach to the study of media attention.