

# **The CMU 2008 Political Blog Corpus**

**Jacob Eisenstein, Eric Xing**

March 2010  
CMU-ML-10-101





# **The CMU 2008 Political Blog Corpus**

**Jacob Eisenstein, Eric Xing**

March 2010  
CMU-ML-10-101

School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

## **Abstract**

This report describes a collection of political blogs on the subject of American politics in the year 2008. The collection was obtained by crawling blog archives in November and December 2009. It is available at <http://sailing.cs.cmu.edu/socialmedia/blog2008.html>.

**Keywords:** datasets, social media

<b>Title</b>	<b>Posts</b>	<b>URL</b>
American Thinker (at)	3197	<a href="http://www.americanthinker.com">http://www.americanthinker.com</a>
Digby (db)	1879	<a href="http://digbysblog.blogspot.com">http://digbysblog.blogspot.com</a>
Hot Air (ha)	3708	<a href="http://hotair.com">http://hotair.com</a>
Michelle Malkin (mm)	677	<a href="http://michallemalkin.com">http://michallemalkin.com</a>
Think Progress (tp)	2080	<a href="http://thinkprogress.org">http://thinkprogress.org</a>
Talking Points Memo (tpm)	1705	<a href="http://tpmelectioncentral.talkingpointsmemo.com/">http://tpmelectioncentral.talkingpointsmemo.com/</a>

Figure 1: Blogs included in the corpus.

## 1 Blogs

The blogs in the corpus are shown in Table 1. They were selected by the following criteria: the Technorati<sup>1</sup> rankings of blog “authority,” ideological balance, coverage for the full year 2008, and ease of access to blog archives.

In the general election for U.S. President in 2008, the following blogs supported Barack Obama: Digby, ThinkProgress, and Talking Points Memo. John McCain was supported by American Thinker, Hot Air, and Michelle Malkin. In general, the blogs that supported Obama in the election tend to advocate for similar policies and candidates as the Democratic party; and the blogs that supported McCain tend to advocate Republican policies and candidates. Digby, Hot Air and Michelle Malkin are single-author blogs; the others have multiple authors.

## 2 Data

The corpus includes all blog posts with more than 200 words (rough word counts were obtained by splitting the text on all whitespace tokens). For each post, there are several files; the shared part of the filename indicates the date of the post, under the format: TTYD00\_NN: TT is a 2 or 3 character abbreviation for the blog title (in parentheses in Table 1); YY is a two digit year marker; DDD is a three digit indicator of the date of the year (from 000 to 365); NN numbers the posts on that day. Note that numbers are skipped when posts are filtered due to having too few words.

There are three suffixes for each post: `.text`, `.xml`, and `.info`. The first suffix is for files that contain the blog text. The text was scraped using blog-specific XPath and regular expressions, designed to exclude boilerplate text, advertisements, and comments. Blockquotes are included.

The `.xml` file contains a subtree of the original document that is guaranteed to include all of the text. Finally, a file with the suffix `.info` contains a set of URLs, one per line: the first is the

<sup>1</sup><http://technorati.com>

URL from which the post was scraped, and the remaining lines show URLs from hyperlinks in the text. Note that there is no guarantee that these URLs are maintained.

### **3 Acknowledgments**

Thanks to Dan Wheeler and Tae Yano for helpful discussions about gathering this data.





**MACHINE LEARNING  
DEPARTMENT**

Carnegie Mellon University  
5000 Forbes Avenue  
Pittsburgh, PA 15213

## **Carnegie Mellon.**

Carnegie Mellon University does not discriminate and Carnegie Mellon University is required not to discriminate in admission, employment, or administration of its programs or activities on the basis of race, color, national origin, sex or handicap in violation of Title VI of the Civil Rights Act of 1964, Title IX of the Educational Amendments of 1972 and Section 504 of the Rehabilitation Act of 1973 or other federal, state, or local laws or executive orders.

In addition, Carnegie Mellon University does not discriminate in admission, employment or administration of its programs on the basis of religion, creed, ancestry, belief, age, veteran status, sexual orientation or in violation of federal, state, or local laws or executive orders. However, in the judgment of the Carnegie Mellon Human Relations Commission, the Department of Defense policy of, "Don't ask, don't tell, don't pursue," excludes openly gay, lesbian and bisexual students from receiving ROTC scholarships or serving in the military. Nevertheless, all ROTC classes at Carnegie Mellon University are available to all students.

Inquiries concerning application of these statements should be directed to the Provost, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, telephone (412) 268-6684 or the Vice President for Enrollment, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, telephone (412) 268-2056

Obtain general information about Carnegie Mellon University by calling (412) 268-2000