An Analysis of Political Views on Blogs
October 20, 2008 Update

Todd Sullivan
todd.sullivan@cs.stanford.edu

1 Introduction

The goal of this project is to analyze political views across the blogosphere. Since the project proposal, we have developed a simple method for extracting feelings about candidates from blog posts and have examined the extracted feelings across a range of attributes such as gender, networks, and time. We have also developed some general ideas for using interests, communities, and friends as additional feeling extraction methods. All analysis up to this point only focuses on feelings towards Obama, McCain, Biden, and Palin.

2 Extracting Feelings from Text

We initially planned to pursue a natural language processing (NLP) approach of parsing each blog post into sentences with part of speech tags and dependency trees, and using this information to accurately extract statements about each candidate from the text. This approach worked well in our past experience with extracting pros and cons from product reviews. Unfortunately, the amount of data in the blog dataset is too large given the timeframe and our current modest computing power (even when only trying to parse text that contains the last name of one of the candidates). Instead, we employ a much more imprecise – but fast – method.

2.1 General Implementation

2.1.1 Preparing the Data

The process starts with tokenizing each blog post into sentences. Aside from separating each post into sentences, this also separates text as needed into words. Text is lowercased, end of sentence markers are separated from the last word in the sentence, as well as commas that may occur midsentence, and compound words are separated. For example, the sentence “I don’t like Fred.” turns into “i do n’t like fred .”. Note that “do” and “n’t” are separated.

After tokenizing the text, we use word lists to compactly define many phrases that are searched for in the text. We first define several word synonyms. Each word synonym contains a list of words that have the same general meaning that we are targeting. We have word lists for bad, good, is, not, plans, I, am, and skip words. For example, the word synonym for “bad” contains idiot, liar, horrible, and boring.

Lists such as “not” and “is” are used as a poor man’s replacement of proper lexical morphology routines that require parsed sentences with part of speech tags. The skip words list is also used as a cheap replacement for our NLP routines that can easily handle extra filler words between the important words in a sentence. The skip words list contains words such as way, so, too, very, and punctuation such as the comma and dash.

We use the word synonyms to define positive and negative lists of feeling indicators. These two lists contain phrases that would indicate a positive or negative feeling about something. For example, “is bad” is one of the phrases in the negative list. Using our word synonyms, and remembering that any number of skip words can occur between words in our phrase, this simple phrase will match phrases such as "is horrible”, “is an

idiot", "is so stupid", etc. To limit the amount of definitions we must maintain, all phrases that contain "good" or "bad" are only defined in the negative list. When reading the lists from disk, we automatically add the opposite of the phrase to the positive list.

Finally, we use candidate name lists that contain many different ways of referring to a particular candidate. This includes obvious choices such as "Obama", "McCain" or "governor of Alaska", and also includes slang /nicknames such as "McSame", "McLame", and "NoBama".

2.1.2 Processing the Data

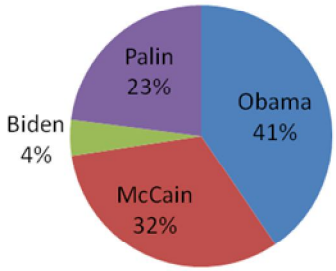
After all of our data is prepared, we extract positive and negative votes from the blog text for each candidate. For each sentence, we identify which candidates are mentioned in the sentence. We then count the number of negative phrases and positive phrases that are present in the sentence and give the counts to each candidate that was mentioned in the sentence. Due to not having the sentence's linguistic parse, we do not make any attempt at anaphora resolution (identifying the entities referred to by "he" or "she").

At the end of the process we have for each candidate the sentences that the candidate was mentioned in and the number of negative and positive statements within the sentences. We use these counts to calculate a FeelScore for each candidate. A FeelScore is a decimal number ranging from -1 to 1 that indicates a feeling with -1 being pure negative, 1 being pure positive, and 0 being neutral. The calculation of this FeelScore is discussed in later sections.

2.2 Sentence Level Analysis

Approximately 6.3% of sentences that mentioned a candidate contains a positive or negative phrase. Figure 2.2a shows each candidate's proportion of blogosphere chatter on a sentence level. The presidential candidates are naturally mentioned in more sentences than the VPs, but Obama seems to be more popular than McCain by 9 points. On the other hand, Palin is talked about 5.75 times as often as Biden. This increased chatter about Palin causes the Republican ticket to have more mentions than the Democrats by 10 points.

Figure 2.2a: Distribution of Candidate Mentions in Sentences

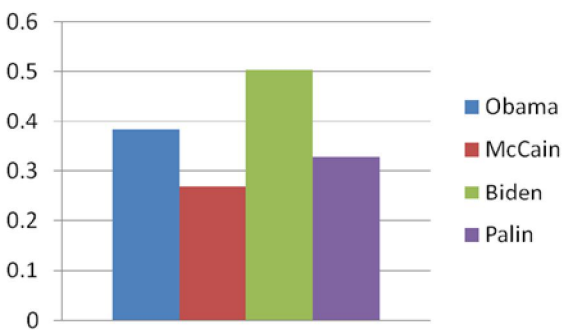


Luckily for the Democrats, not all publicity is good publicity. Figure 2.2b shows the sentence-aggregate FeelScore for each candidate. We calculated the sentence-aggregate FeelScore by summing the positive and negative counts for each candidate across all of the sentences. We treat each positive or negative count as a vote and use the following equation to calculate the candidate's FeelScore:

$$\frac{PositiveCount - NegativeCount}{PositiveCount + NegativeCount}$$

This results in a number in the range [-1, 1].

Figure 2.2b: Overall Candidate FeelScores (Sentence Level)



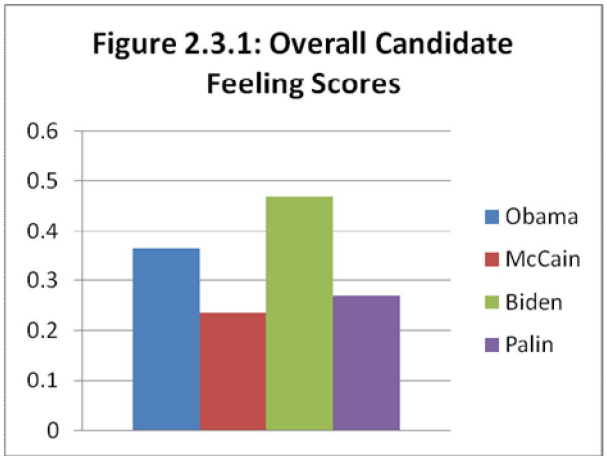
As Figure 2.2b shows, all candidates have an overall positive FeelScore. Despite having a small presence in blogs, Biden has the highest FeelScore. When comparing the Republican and Democratic tickets in terms of FeelScore, the Democrats are clearly the winners. Somewhat interestingly, both vice presidential candidates have a higher FeelScore than their respective runningmate!

2.3 Author Level Analysis

The sentence-level FeelScore is inadequate because a single spammer can overly manipulate the scores by posting a single blog message that contains thousands of positive or negative mentions for any of the candidates. To mitigate this issue, we calculate a FeelScore for each author/candidate pair in the same way as before. We then calculate an overall FeelScore for each candidate by summing each author's FeelScore for the candidate and dividing by the number of authors that expressed a positive or negative view about the candidate. This process produces a number in the range [-1, 1].

2.3.1 Overall Candidate FeelScores

Figure 2.3.1 shows the overall FeelScores for each candidate. Comparing the result with the sentence-level FeelScores in Figure 2.2b shows that while all FeelScores dropped a little, the overall shape of the scores relative to each other did not change.



2.3.2 FeelScores by Blog Network

We found that the FeelScores vary significantly by blog network. Figure 2.3.2a shows the positive/negative chatter proportions for each candidate on each network. LiveJournal and Blogger have almost identical distributions while chatter on MySpace seems to be all about the presidential candidates – and mostly about Obama.

Figure 2.3.2a: Positive/Negative Chatter By Blog Network

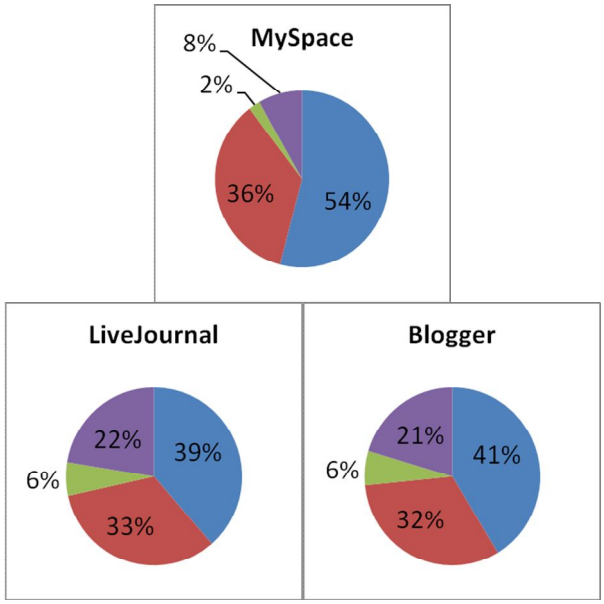
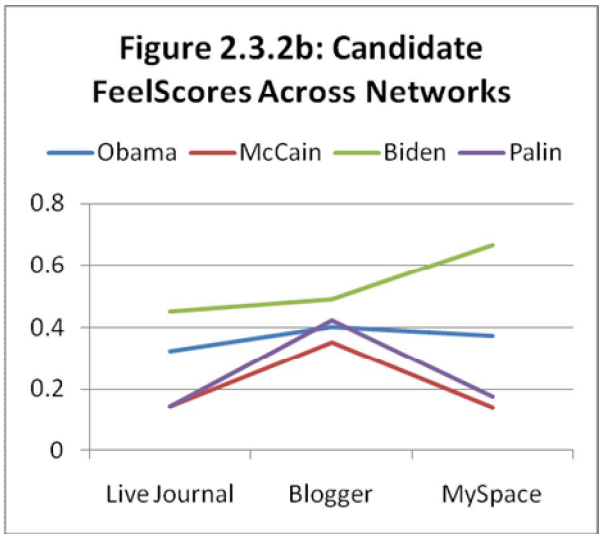


Figure 2.3.2b shows the FeelScores of each candidate across each blog network. The data shows that the increased proportion of chatter about Obama on MySpace is from roughly equal levels of positive and negative remarks since Obama's FeelScore remains fairly constant across the networks. None of the other candidates have as flat a FeelScore curve across networks as Obama.

Biden remains with the highest FeelScore regardless of the network. The Republican ticket appears to have a stronger base on Blogger than on LiveJournal, with both McCain and Palin exhibiting the same curve that is much higher at the Blogger network than at LiveJournal or MySpace. Biden's jump in FeelScore at MySpace is most likely

due to only being mentioned 6 times – the data is scarce and unreliable.



2.3.3 FeelScores by Gender

Much of our data contains gender information about the authors. Each author either has a gender of unspecified, female, male, or X. At the time of this writing we do not actually know what X stands for, but we will have that answer by our next update.

Figure 2.3.3a shows the positive/negative chatter of each candidate based on the gender of the author. Both the unspecified and X genders have the same proportions, which happen to be the same proportions of the Blogger and LiveJournal networks and close to the distribution of candidate mentions in sentences (Figure 2.2a). Authors that specified themselves as male or female talked more about the presidential candidates, which could be because gender information is more abundant on MySpace and MySpace was shown to have more presidential chatter than normal.

Males seem to have less positive/negative comments to say about Palin, and slightly more than everyone else in regards to Biden. On the presidential side, males talk much more about McCain than females do. Females predominately talk about Obama while almost not talking at all about Biden. Compared to males, females make more comments about Palin.

Figure 2.3.3a: Positive/Negative Chatter By Gender

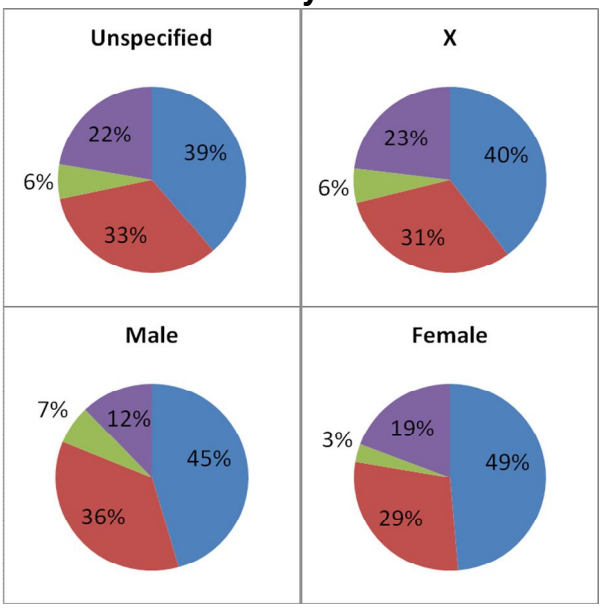
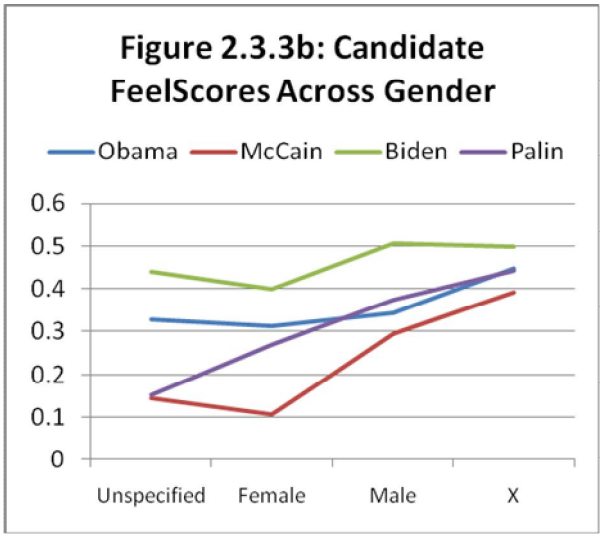


Figure 2.3.3b shows each candidate's FeelScore across gender. Similar to the networks case, Obama's score seems to stay more level across gender than the other candidates. Females are the most negative group towards Biden and McCain, which may or may not be attributed to the two candidate's older ages in comparison with Obama and Palin. The figure shows that the Republican ticket is most competitive against the Democrats in the male and X groups while being far behind in the female and unspecified groups.



3 What People Are Saying

As an addendum, here are several sample phrases showing what people are saying about the candidates.

```
Obama for Prez

Both have huge flaws, but on balance McCain's is better

The problem with McCain's approach…

I disagree with Obama on the war…

Right away, Biden loses focus on the actual topic

Someone like Joe Biden or Chris Dodd would be good

As second choices go, Biden is a good one

I can think of none better than the current Governor of the good state of Alaska
```

4 Future Work

While we did not prepare graphs and analysis for FeelScores across time, we have data calculated. In the coming weeks we will pursue many avenues with the blog data, such as the FeelScores separated by other dimensions as well as the change in FeelScores across time. We will compare the time-based FeelScores with important events such as debates, signing of the Wall Street bailout package, and any large blunders such as Palin's Couric interviews.

We will also incorporate additional information into calculating the FeelScore. This includes using the same method of finding feelings from text but targeting generic terms such as "Republican" or "Democrat", as well as targeting issues such as gun control, taxes, and healthcare. We will also pull information from the interests, communities, and friends of the authors in an effort to increase coverage.