LING 570 Project 1

Jennifer Romelfanger and Chris Curtis

6 Nov 2012

Summary

For our classifier we chose to take a largely different path than a majority of the class. Instead of writing a separate program to build the feature vectors, we extended the Mallet Java API to add our desired features to the pipeline. This eliminated the issues with the import files other student had, but introduced a separate layer of confusion in dealing with poorly documented code, and bizarre mapping structures.

What features/words were the strongest indicators?

On the right side, we found that proper names became the strongest indicators. This includes words like "Fox", "Congressman", "Mr. Obama", and other names. It was very interesting that "Mr. Obama' showed up at all, but this could represent a form of address chosen deliberately to show a lack of respect for the president.

The left side had a stronger lean towards more common words such as "explain", "army", and "exactly". The left also includes more terms such as "sex," "gotta," and "passover," which would be rarer in the Conservative Christian narrative style that the right appears to prefer.

What features beyond words did you use? Why? What difference did each of the features have on your classifier?

We included bigrams that had a count greater than 3 in the training data. Prior to counting these bigrams, we also removed stop words from the text so that the bigrams would only include relevant terms. We chose this as an option that could help capture lengthier topics, such as "health care", "President Obama", and "marriage equality". Adding the bigrams increased the accuracy of the classifier on the devtest and test data, and helped balance the incorrect identifications on each side.[1]

We added a step in the pipeline that converted all numbers to ##. This allowed us to keep the concept of numbers without any particular number being picked out. We decided to add this because "12" was showing up as a strong indicator for right, and it seemed likely that this could improperly affect general cases. This change did not have any obvious affects in our tests, but we kept it due to its potential to help with generalization.

Another step was added to remove HTML. This helps with generalization and removes text that is not relevant to the content of the article[2]. This mildly reduced our accuracy, though, because HTML tags were not

[1] Initially there was a lot of incorrect identifications on the left side.

[2] Such as JavaScript code!

uniformly distributed in the training data and so ended up being relatively strong predictors when included.

We added a Porter Stemmer in order to group terms like "Republican" and "Republicans" together. This added to our accuracy and helped make the classifier more stable. This also made it more difficult to find features that would significantly affect the output, as it collapsed any fine distinctions in word choice (for example, "Democratic Party" vs. "Democrat[3] Party").

We also added a Flesch-Kincaid reading level score as a feature. A higher reading level was slightly predictive of left, and empirical testing led us to add a binary feature for reading level > 9.0, which was more strongly predictive (though still a minor predictor).

One additional extra feature we added was the ratio of the use of the term "liberal"/"conservative". This slightly helped with our accuracy.

One feature that we considered using was punctuation. This feature significantly increased our accuracy during tests (from around 92double quotes. We decided to remove this feature as it did not generalize well and had the potential to strongly skew our media test results.

[3] This is a common usage on the political right that deliberately misstates the name of the party.

What features would you add if you could?

Quotes seemed to be our biggest issue. The left side often contained quotes that contained words that should have been a strong indicator for a right-leaning article. We would have liked to add a feature that could differentiate quoted text from the main text of the article.[4]

Source metadata would be an interesting addition (including, perhaps, authorship) but we suspect it would be far too strongly fit to known data—authors who write exclusively for one side would be overly strong indicators of bias.

[4] Apart from the computational complexity of identifying quoted material, there may be additional confounding factors, such as whether the quotation represents approval or disapproval.

Where do you think your features are weak?

We think our features still pick up on names and words that are relevant only in the time period of the articles.[5] If the articles are used to identify text from another time period, the classifier will likely be less effective. Also our features poorly handle quotes, as stated above.

We also suspect our features are weak in their cross-domain applicability. That is, there is a reasonable argument to be made that the training domain is not coextensive with the news media domain, and so the trained features are perhaps not as applicable as we would like. [6]

[5] "Healthcare reform" means something entirely different today than it did six years ago.

[6] For example, using a classifier trained on a WSJ corpus to analyze articles from Fast Company magazine would present similar concerns.

Do you believe your classifier? What did it capture? What did it fail to capture?

Our classifier captured that media sources are likely to be right-leaning compared to Mother Jones. It also captures how largely different the vocabulary of each side is. It's not entirely clear, however, that the Fox data is equivalently polarized, given its (at least tenuous) connection to the "mainstream" news site.[7] This has the effect of generally shifting the center of mass towards the (theoretical) right.

[7] As compared to, say, National Review Online or The Free Republic.

Out classifier also didn't effectively capture sentiment. This is most evident in the inability to deal with quoted material appropriately, but the primarily word-driven features simply don't give enough specificity to address the complex senses in which phrases are cited and used. The inability to properly capture quotations negated the usefulness of deliberate partisan terminology choices.

One interesting result is the difference between the left and right score rankings. For example, the CSM source material had the highest average left *and* right scores. This could be consistent with multiple diverse but polarized authorial voices.

In general, though, the classifier seems to produce a reasonable index of media "bias" that seems to correlate with real-world intuition.