Predicting Newspaper Political Leanings from Twitter Followers

The Party Guys: jgraves1, ksachan, nmercha2, jzhu71

Goal

As America's largest news sources become increasingly polarized, it's important to be aware of their political biases in order to responsibly inform ourselves. While editors can assign partisanship scores to newspapers based on their articles, the process is neither objective nor scalable. There's also limited/no data on the biases of local papers.

We aim to predict the political leaning of news outlets based on the political leanings of their Twitter followers, which we determine by looking at the politicians those users follow. This model should hopefully scale to all news organizations with a Twitter account.

Data

We collected Twitter data on 180 news organizations, including national outlets such as Fox/NYTimes and local outlets such as the Providence Journal and the Brown Daily Herald. We collected 200 followers of each paper and gave each follower a political score based on the politicians they follow. We used three different external datasets to score politicians. To validate our prediction of the political leaning of a newspaper, we used a dataset from All Sides that scored 55 of the newspapers we looked at.

Our scores were skewed left (the average score given by our algorithm to a Twitter user was -0.318, with negative meaning more liberal). This does not mean that our algorithm is flawed because this score is dependent on the news outlets we are sampling followers from. Also, studies show that Twitter users are more liberal than the average American.

~30% of the users we looked at followed between one and three politicians. Most of those users follow Trump. This significantly reduces our predictive power because many people who follow Trump on Twitter aren't conservative.

Model+Evaluation Setup

Each newspaper was given a distribution that was the histogram of the political leaning of its 200 followers. To address the Trump effect, we smoothed the data. We tested both weighting each follower by a function of the number of politicians they follow and adding three zeros to their score before averaging. Both techniques also had the positive result of giving a larger weighting to users who follow more political accounts. This makes sense because we can be more confident in the political leanings of those users.

Since we wanted to classify newspapers based on their political leaning, a classification mechanism made sense. We initially tried using a linear regression, since our labels had a natural ordering with (1) being liberal, and (5) being conservative. We used K-fold validation to account for our small dataset, and at each iteration recorded the goodness-of-fit (R-squared) and prediction mean squared error for the given train and test split. On testing data, our MSE would often range higher than one. Our best guess is this reflects the fact that even in the case where our model can assign an intermediate label, for example 2.5, there is significant confusion between center and left-leaning, and center and right-leaning papers.

Giving the middling success of our linear regression, we tried an alternative approach. We decided to use a decision tree since many of our features on the newspapers distribution may not influence the classification in a linear way. To ensure we didn't overfit the data, we used a bootstrap method, breaking the data into a train test split and iterating 20,000 times. To improve the model we experimented with feature selection and optimizing the maximum depth of the tree.

Results and Analysis

Claim #1: We can accurately predict the political leaning (the bias in the articles that they are writing) of a newspaper by examining who 200 of their followers follow.
Support for Claim #1: Both the linear regression and decision tree models performed well, explaining a large portion of the variation in political leaning, and accurately labeling the newspapers over 60% of the time, without noticeable overfitting.

Model	Train Data Results	Test Data Results
Linear Regression	0.8 (R^2)	0.69 (R^2)
Decision Tree, 5 Class	63% (Accuracy)	63% (Accuracy)
Decision Tree, 3 Class*	87% (Accuracy)	77% (Accuracy)

*Our 3 Classes in this case are created by aggregating left and center left, and right and center right

Claim #2: The Decision Tree did considerably better using feature selection, and optimizing the depth of the tree. Our model is generally correct a much higher percentage of the time.
Support for Claim #2: As the confusion matrix shows, the model is generally accurate a much higher percentage of the time. The model basically never confuses right/lean right papers with left/lean left papers and the most common errors are misclassifying a paper as center when it is not.

All Sides \ Predicted	Left	Lean Left	Center	Lean Right	Right
Left	10400	6892	4704	0	3
Lean Left	8810	13168	4270	72	31
Center	4278	5889	17657	8	311
Lean Right	0	9	7	15571	6022
Right	1	22	2224	4319	15332

Confusion Matrix for the decision tree test data set after bootstrapping and running 20,000 iterations

Claim #3: There are features (besides mean score) of the distribution of a papers followers that are useful in predicting its All-Sides bias rating
Support for Claim #3: Running a linear regression presented some difficulties- namely, how do you take a list of scored users, essentially 200 numbers representing samples from some distribution- and extract any sort of meaningful features to regress against? We tried a number of possible 'embedding' strategies, and while we were unable to find something we were totally satisfied with, many of them allowed for good or at least non-trivial prediction using linear regression. Because of this, we believe that a more intelligent procedure for embedding our 200 distribution samples into some kind of vector space may give us an even stronger linear correlation with the All-Sides bias ratings.

Claim #4: There is reason to believe that our model scales well to local papers.
Support for Claim #4: One limitation of the All Sides data is that it doesn't include local papers. However our model had some of the most difficult time classifying large newspapers like the New York Times, giving us reason to believe that it would work well for smaller newspapers. If this classification does indeed generalize well to local newspapers, we use it to quantify political bias for a broad and frequently overlooked portion of the national media market.

Data Sources

- All Sides Dataset: https://www.allsides.com/media-bias/media-bias-ratings#ratings
- Scored Politicians and Top 200 Political Accounts Dataset: http://pablobarbera.com/static/barbera_twitter_ideal_points.pdf
- Official Campaign Accounts: https://www.propublica.org/datastore/dataset/politicians-tracked-by-politwoops