Predicting Newspaper Political Leanings From Twitter Followers

John Graves Kiran Merchant Jeffrey Zhu Kshitij Sachan

BROWN Computer Science

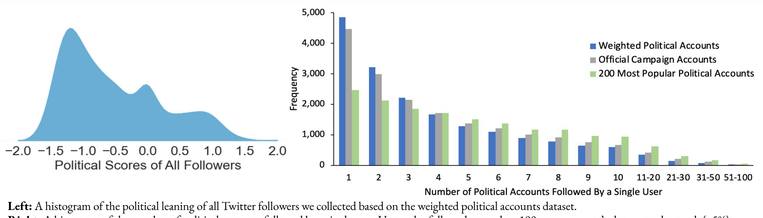
Introduction

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. Our project aims to model and predict the partisanship of newspapers using the political leanings of their Twitter followers. By learning the relationship between a newspaper's bias and the leaning of its readership, we can provide objective scores to large news sources, as well as scale to smaller sources and provide transparency in local newspapers.

Data Collection

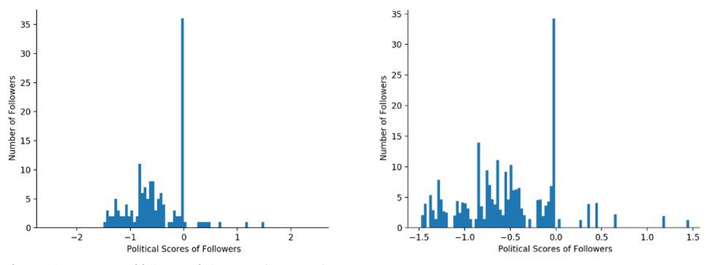
- We collected **Twitter** data on **180 news organizations**, including national outlets such as Fox/NYTimes and local outlets such as the Providence Journal and the Brown Daily Herald
- Twitter Pipeline:
  1. Collected 900,000 most recent followers of each paper. Filtered to remove companies/bots. Removed account if it followed less than 25 people or the name included top 1,000 most common words, numbers, special characters, etc.
  2. Randomly selected **200 followers per paper** and collected the 5,000 accounts they followed most recently.
  3. Gave each follower a political score based on the politicians they followed; the distribution of a paper is simply the scores of its 200 followers
- Political dataset - We used three different datasets to score politicians
  - Weighted political accounts - Only 600 politicians but continuous scores. Dataset is from 2016, so slightly dated (e.g. Trump has a score very close to 0)
  - Official campaign accounts - Given a binary -1/+1 score by us
  - 200 most popular political accounts - From research paper; also has binary +1/-1 score
- Prediction Validation: 55 of the papers we looked at were scored by All-Sides on a 1 to 5 scale

Data Observations



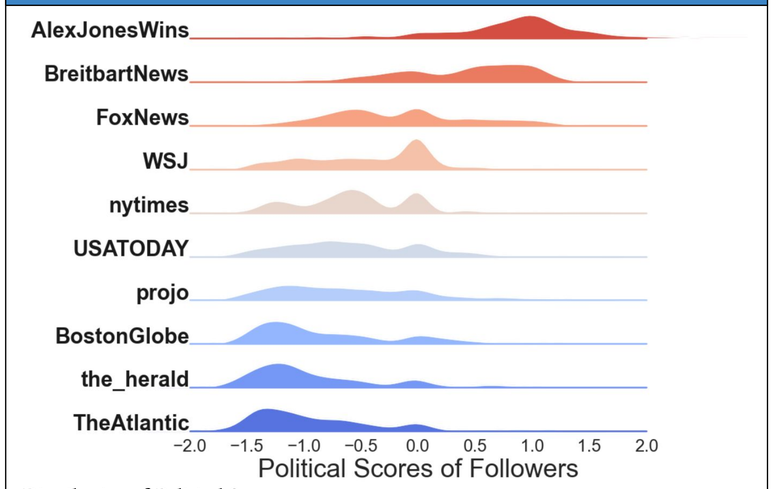
Left: A histogram of the political leaning of all Twitter followers we collected based on the weighted political accounts dataset.
Right: A histogram of the number of political accounts followed by a single user. Users who followed more than 100 accounts aren't shown on the graph (~5%).

1. Top right: ~30% of users followed <=3 political accounts. This limits our predictive power because most of these people follow Trump/Obama.
2. Top left: Distribution of political users was skewed left (mean = -0.318), consistent with studies showing the average Twitter user is more liberal than the average American.
3. Bottom left: The Trump Effect
   a. The spike near 0 is caused by the people who only follow Trump (who has a score near 0 according to our 2016 weighted political accounts dataset)
   b. Most people follow Trump because he is the president, not because they are conservative
4. Bottom right: Adjusting for the Trump Effect and people who follow few politicians
   a. Especially harmful when each politician has a +/-1 score like in the second two datasets
   b. Someone can follow one account and have the same score as someone who follows many
   c. We attempted to smooth our data in two ways:
      1. Weight each user by the cube root of the number of political accounts they follow
      2. Add three zero scores to each person's political score before averaging



Left: Initial distribution of followers of Bloomberg (@Business)
Right: Distribution after weighting users based on number of accounts followed using a cube root

Visualizing Distributions of Follower Scores



Distribution of Political Scores

- Above: Distribution over the political scores of followers for each newspaper using kernel density estimation
- The mean, variance, etc. can be used as features to estimate a newspaper's political leaning

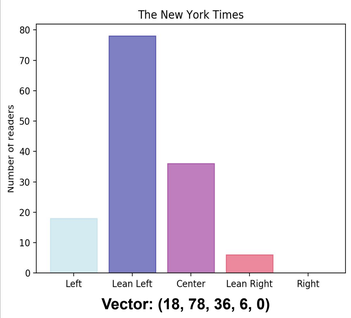
Embedding & Linear Regression

Given a list of the political scores of readers, sampled from some true underlying distribution for that paper (like the ones approximated above), how do we extract a feature vector useful for linear regression? Two approaches we tried:

1. Create 5 equal length bins, ranging from -2 to 2, sort the political scores into these bins, and return a 5-d vector of the resulting bin counts.
2. Run a 2-cluster K-means on the list of political scores. Calculate the mean and variance of each cluster, as well as the overall mean and variance, and return a 6-d vector of these values

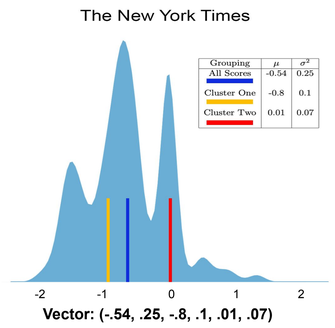
Visualize how each approach embeds, for example, the distribution of The New York Times

(1) Naive Binning



Vector: (18, 78, 36, 6, 0)

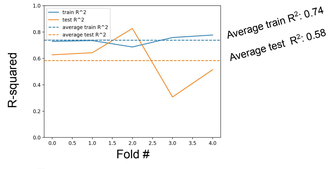
(2) Cluster Summaries



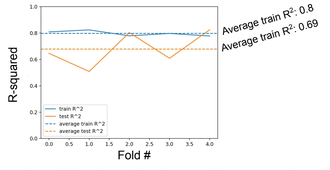
Vector: (-.54, .25, -.8, .1, .01, .07)

To measure the usefulness of each embedding, we ran a K-fold validation, at each step training a linear regression on the training vectors against their All-Sides bias score and calculating a goodness-of-fit (R-squared) on both the training and testing vectors.

(1) Naive Binning



(2) Cluster Summaries



- Both vector embeddings were high-dimensional, and both suffered significant overfitting
- Our cluster summary embedding was designed to capture useful information about the whether a distribution was one or two peaked, and where those peaks were.

Decision Tree

Machine Learning Decision Tree

- Predicted political leaning of papers
- Used train/test split with bootstrapping (~20,000 iterations) to address overfitting
- Tested many different sets of features including filtering the number of politicians followed and averaging over multiple parameters

Features	Max Depth	5 Categories	3 Categories
Version 1	4	60.10%	82.80%
Version 2	3	63.90%	76.50%
All	3	49.90%	74.20%

Accuracy of different decision trees

Feature Selection

- Found two sets of features with best accuracy:
  - Smoothed mean and unsmoothed variation on weighted political accounts
  - Unsmoothed mean and semi-smoothed variation on top 200 accounts

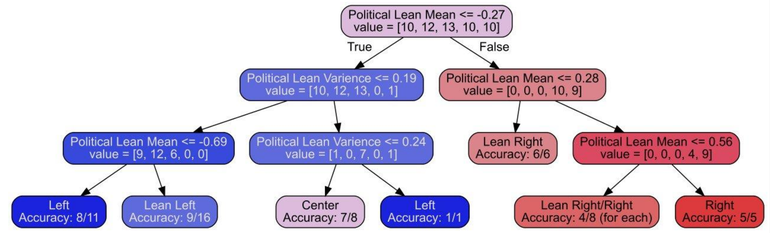
All Sides \ Predicted	Left	Lean Left	Center	Lean Right	Right
Left	10400	6892	4704	0	3
Lean Left	8810	13168	4270	72	31
Center	4278	5889	17657	8	311
Lean Right	0	9	7	15571	6022
Right	1	22	2224	4319	15332

All Sides \ Predicted	Left	Lean Left	Center	Lean Right	Right
Left	14330	236	7060	0	2
Lean Left	3989	9079	13030	0	0
Center	3018	634	22730	4	2022
Lean Right	0	0	226	17431	4266
Right	0	0	2236	6625	13082

Confusion Matrices describing the distribution of prediction errors from decision trees: **Left:** Version 1 **Right:** Version 2

Our Model

- Selected Version 2 of Model
- Compared to Version 1:
  - Predicted All Sides Data with a higher Accuracy
  - Lower accuracy three political categories instead of five
  - Had less overfitting (Smaller depth and smaller gap in train test split by ~0.16)



Final Decision Tree Model using Mean and Variance. Trend looks generally correct with slight overfitting, especially for the Left (1/1).

Ethical Considerations

- Our primary ethical consideration was protecting the anonymity of readers. Our pipeline both collects the twitter handle of, and calculates a political score for, nearly 25,000 twitter users.
- While our analysis is done in aggregate, and therefore anonymized, there are many intermediate steps in our pipeline where personally identifying information is stored, for example, in plain text or CSV files.
- Files with personally identifiable information were only ever visible to us or the TA staff, and when our analysis is complete we will delete or otherwise anonymize any remaining data.

Limitations/Further Work

- Twitter Limitations
  - Capped by the Twitter API at 180 users/hour
  - The average person on Twitter is very different from the average person in real life
  - Most local papers don't have Twitter followers or have very few Twitter followers
  - Much larger Facebook dataset is restricted to researchers
- Further work
  - Explore other models for classification (e.g. logistic regression)
  - Generate better vector embeddings of our distributions, maybe using Expectation Maximization to model each distribution as a mixture of Gaussians
  - Find baseline bias labels for more papers, especially local papers with less readership