Ideology and News Content in Contested U.S. House Primaries*

Andrew B. Hall†
Department of Political Science
Stanford University

Chloe Lim‡
Department of Political Science
Stanford University

July 9, 2018

Abstract

Pundits and scholars often claim that congressional primary elections favor extremist candidates, but the mechanisms by which primary voters might learn about candidate platforms are not well understood. In this paper, we collect a new dataset of roughly 16,000 local newspaper articles matched to candidates in contested U.S. House primary races from 1998 to 2012. Using supervised machine learning, we classify these articles into political topics. On average, we find little coverage of candidate platforms. However, we also find that the advantage of extremist candidates in House primaries—measured using the campaign contributions they receive—is concentrated in elections with low levels of newspaper coverage. Where newspaper coverage is higher, there is more coverage of candidate platforms, and extremist candidates do worse. The results suggest that the advantage of more extreme candidates in contested House primaries may be the result of information failures and not just the preferences of primary electorates, and that extremist candidates may do increasingly well as local newspaper coverage continues to decline.

*Authors contributed equally and are listed in alphabetical order. For comments and suggestions the authors thank Justin Grimmer, Shigeo Hirano, and Jim Snyder. For research assistance, the authors thank Elise Kostial. For data, the authors thank Shigeo Hirano and Jim Snyder.

†Andrew B. Hall is an Assistant Professor in the Department of Political Science at Stanford University (`http://www.andrewbenjaminhall.com`).

‡Chloe Lim is a Ph.D. Candidate in the Department of Political Science at Stanford University (`http://www.chloelim.org/`).

1 Introduction

Polarization in U.S. legislatures is at all-time highs (e.g., McCarty, Poole, and Rosenthal 2006), leading scholars and pundits to search for its roots in electoral politics. A popular claim is that congressional primary electorates prefer more extreme candidates, which in turn causes legislative polarization (see, e.g., Owen and Grofman 2006; Pildes 2011, but also see Boatright 2013; Hirano et al. 2010; McGhee et al. 2014). A policy brief from Brookings, for example, writes that "The electorates in [primaries] tend to be small...and often unrepresentative. Hence, candidates are frequently forced to protect their flanks by moving away from the center."[1] Existing research does suggest that congressional primary electorates prefer more extreme candidates, on average (Brady, Han, and Pope 2007; Hall and Snyder 2015*a*; Thomsen 2018), but the magnitude of the advantage is modest. At the same time, it is still an open question whether primary electorates are unrepresentative of voters more generally or not (Hill and Tausanovitch 2017; Sides et al. 2017). Whatever the truth, a missing link in any account of whether or how primary electorates support more extreme candidates is information; if primary voters intentionally support more extreme candidates because of their platforms, then at least some pivotal subset of voters must have information about the platforms of primary candidates. Little research has directly investigated what information is available to primary voters in real elections.

To accomplish this goal, we collect a new dataset from online sources containing the headlines and summaries of approximately 16,000 local newspaper articles about primary candidates in U.S. House races over the time period 1998 to 2012. Based on a careful reading of a random sample of several thousand of these headlines and summaries, we use supervised machine learning to sort the articles into six mutually exclusive categories. After validating the classifications using third-party hand codings and correlations with real-world events, we find that, on average, local news provides voters with little information about their primary candidates' platforms. The average candidate in a contested House primary

[1] `https://www.brookings.edu/research/thinking-about-political-polarization/`

is mentioned in only 3.3 articles in total, and we estimate that 75% of these articles provide what we call basic campaign coverage—articles about the bare facts of the race, like who is running and who has dropped out, that convey no ideological information at all. More than three quarters of the races in our sample appear to have no news articles covering candidates' stated position at all, while another three quarters have no articles about endorsements, and more than half have no articles about platforms or endorsements, combined.

This does not mean that news coverage is not informative, however. We also show that the share of news coverage in a primary race is a useful predictor of who will win the race. Even if primary news coverage is often short on details, newsworthiness itself is an important leading indicator of electoral outcomes.

Finally, we show that newspaper coverage appears to help more moderate candidates—contrary to the claim that primary voters prefer more extreme candidates. Using estimates of candidate ideology estimated from campaign contributions, we find that primary electorates' preference for more extreme candidates is concentrated in low news-coverage areas. Where newspaper coverage of the primary election is higher, more moderate candidates receive higher average vote shares than where it is lower. Although our measure of newspaper coverage is not randomly assigned, we do what we can to suggest that the relationship is causal by using a potentially exogenous measure of local news coverage based on congruence between a newspaper's circulation and the local congressional district (Snyder and Stromberg 2010).

Adding to the plausibility of these analyses, we find that, where congruence is higher, there are more news articles are about candidate platforms. This suggests that primary voters may support more extreme candidates less when they learn more about candidate positions. It is difficult to square these results with the claim that primary voters genuinely prefer more extreme candidates; if that claim were true, more news coverage of candidate positions should help primary voters to pick out more extreme nominees.

In addition to its relevance for the study of primary elections and polarization, our paper also adds to the literature on voters and information in two main ways. First, by examining news coverage directly, we are able to evaluate the actual information environment that primary electorates face. This is in contrast to survey studies that manipulate the information environment, but at the risk of not reflecting the real-world information environment (e.g., Fowler and Margolis 2014; Riggle 1992). The patterns of news coverage that we document may be useful for future survey-based studies that wish to emulate the real news environment. Second, we are able to study the key mechanism by which aggregate patterns of news coverage—investigated in, for example, Peterson (2017) and Snyder and Stromberg (2010)—actually influence public opinion and electoral choices. The fact that higher congruence areas also see more coverage of candidate platforms suggests that media coverage may influence voter behavior directly. In addition, the fact that news worthiness is itself a predictor of candidate performance suggests that newspaper coverage may help primary voters to vote strategically and avoid wasting votes (e.g., Hall and Snyder 2015*b*).

Taken together, our evidence suggests that the continued decline of local news organizations—documented, for example, in Peterson (2018)—will increase the tendency to nominate more extreme candidates, unless alternative sources of information substitute for the campaign coverage that local news has historically provided. This is consistent with evidence concerning local television news (Gilliam and Iyengar 2000; Martin and McCrain 2018).

2 What Can Primary Voters Learn From News Coverage?

In this section, we evaluate the content of local newspaper headlines and article summaries related to contested U.S. House primary elections.

2.1 New Data on Newspaper Articles About House Primaries

We collected information on primary news articles from `NewsLibrary.com`, an online archive of articles published by over 6,000 newspapers from across the United States. It has a wide variety of local newspapers of varying circulations, ranging from large-scale newspapers like *The Sacramento Bee* and *New York Daily News* to small-town daily newspapers such as the *Eaglewood Sun*. `NewsLibrary.com` was used in Snyder and Stromberg (2010) to collect local news articles about members of the U.S. House.

To query NewsLibrary, we start from a dataset on U.S. House primary elections, originally collected by Ansolabehere et al. (2010) and extended to subsequent years by the same authors. Along with a full range of electoral variables (like vote share), the dataset contains the full name of each candidate. Using NewsLibrary, we collected the headlines and summaries for local newspaper articles for these candidates in all contested U.S. House primary elections from 1998 to 2012.[2] Each search was confined to newspapers that were published in the state in which the candidate ran between January 1st and the primary election date for each election year and state, using the following search terms: [Candidate's Last Name] in Headlines, ["Primary"] in All Text, and ["Candidate"] in All Text. We deleted articles that were published on or after the primary election date (which varies across states) for each corresponding state, because we are focused primarily on information that is available to voters prior to the election day.[3]

To clean the resulting dataset, we manually scrutinized articles for candidates whose last names are among the 100 most common in the 2010 census, to determine whether each article was in fact about a House primary candidate. We likewise manually checked articles where the lead paragraph did not include the first name of the candidate. By reading all of these articles, we excluded those written about a different person with the same last name running for different office (e.g., County Commissioner, District Attorney, etc.) or about the

[2]The archive does not grant open access to full article content, but it does allow us to view headlines and summaries that provide relatively detailed information about the article.
[3]We obtained data on primary election dates from the FEC website.

same person running for another office in the same election year, after having resigned from the House race.

We also use the FEC IDs from the election dataset to merge candidates with their ideological scalings originally developed in Hall and Snyder (2015*a*), and extended to later years in Hall and Thompson (2018). These scalings impute ideological positions for candidates who've never held office in a two-step process. First, donors are scaled based on the DW-NOMINATE scores of incumbents that they donate to—so, for example, a donor who donates to candidates who have far-right Nominate scores is imputed to be a far-right donor. Second, candidates are scaled based on the donors from whom they received contributions—so, for example, a candidate who receives donations from donors that support far-right incumbents is estimated to be a far-right candidate. The resulting scalings correlate well with DW-NOMINATE, even within-party. For further discussion of the validity of these scalings, see Hall and Thompson (2018).

The final dataset includes 2,448 candidates and 15,801 articles published in 2,039 local newspapers.

2.2 Classifying Primary News Coverage

Our goal is to understand the content of newspaper coverage about U.S. House primaries, and to evaluate whether it offers significant information about candidate platforms. Before applying any methods, we read 5,000 article headlines and summaries ourselves. Based on our reading, we defined six categories of news coverage: Campaign Coverage, Endorsements, Candidate Biographies, Money, Platform, and Scandal. The categories are largely self explanatory. Campaign Coverage articles discuss the bare facts of the race, such as who is running and who has dropped out. Articles in the Endorsements category report endorsements that candidates have received. Articles in the Candidate Biographies category provide specific information on candidates' backgrounds, like their professions, any previous political offices they have held, their age, and so forth. Articles in the Money category are focused

on information about candidate fundraising. Articles in the Platform category, which we are particularly interested in, report specific policy positions or ideological views that candidates have offered. Finally, articles in the Scandal category focus on potential scandals related to a candidate in the race. Because we do not have access to full article content, it is possible that an article classified in one category based on its headline and summary could contain paragraphs that would fit into other categories; however, when we compared full content for articles we were able to find on other websites online, we rarely found this to be the case (most articles are quite short, and the summaries generally indicate the full scope of their content.)

Appendix A offers more details on the categorization scheme used to classify the news articles, and offers specific examples of articles coded into each category. We automatically coded any article including the word stem "endors" as Endorsement and assigned 4,828 articles to one of the remaining five categories manually.

We then used these 4,828 researcher-coded entries as a training set for a variety of supervised machine learning procedures (see for example Grimmer and Stewart 2013; Lucas et al. 2015). Specifically, we compared the performance of eight different classifiers: Support Vector Machine, Maximum Entropy, Supervised Latent Dirichlet Allocation, Boosting, Bagging, Random Forest, Neural Network, and Tree (we did not include the Endorsement category articles since their coding is deterministic.) Of the 4,828 entries, 2,414 were randomly chosen to train each model. We measured how well each model performed on the entries that were not used to train the model by comparing the classification results against the researcher codings.

Table A.1 shows precision, recall, and f-scores for each algorithm. In the context of our research, precision is the proportion of news articles correctly classified as Category A out of the total number of articles classified as Category A by the algorithm. Recall is the proportion of news articles correctly classified as Category A out of all true Category A articles (i.e., all articles assigned to Category A by the researcher). Recall is a function of both true positives

(Category A articles correctly assigned to Category A) and false negatives (Category A articles incorrectly assigned to other categories). F-scores are a weighted average of both precision and recall.

We choose to focus on SVM, since it has the highest F-score, at 0.734. Having chosen SVM, we trained it on the entire training set of researcher-coded entries. We then applied it to the rest of the newspaper articles which were not classified by the researcher, providing us with topic classifications for each article in our sample.

2.3 Validating Our Primary Election Newspaper Coverage Classifier

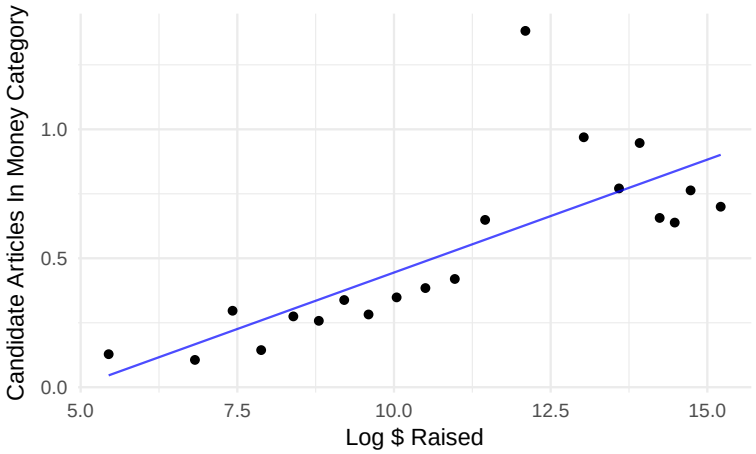
Before using the SVM classifications to analyze news content, we validate our approach in several ways. First, and most importantly, we asked an independent coder to categorize a random sample of 1,000 articles from the test set. The intercoder reliability rate between our codings and the codings of the independent coder was 0.77, meeting the standard of intercoder reliability that academics commonly apply when evaluating hand-coded data (Barrett 2001).

Second, we correlate the amount of money each candidate in our dataset raises with the number of articles about that candidate that are classified in the 'Money' category. Figure 1 presents a binscatter showing the strong correlation between the two variables. Candidates who raise more money also have more articles classified into the Money category, suggesting that, at the very least, the Money topic categorization is meaningful.

Third, we investigate which specific words are most predictive of each categorization. To do so, we create a document-term matrix (DTM) from the article summaries, and we combine this matrix with the SVM categorizations for each article.[4] We construct a dummy variable for whether each article is a member of each of the six categories, and we run ridge regressions

[4]We construct the DTM using the create_matrix() function in the RTextTools package, using options to remove numbers, stem words, and remove stopwords. We set the removeSparseTerms option to 0.99.

Figure 1 – Validating Article Classification Using Fundraising Data. Candidates who raise more money have more articles classified in the 'Money' Category.



Note: Points are averages in equal-sample-sized bins of Log $ Raised. Regression line fit to underlying data. Generated using `binscatter` in Stata.

Table 1 – Informative Words by Category. Presents the five words most predictive of an article being classified as each of the six mutually exclusive topics.

Campaign Info	debate,district,candidate,percent,rep
Bio	series,education,university,candidate,college
Endorsement	endorsed,endorsement,endorsements,endorse,endorses
Money	money,raised,fundraising,fund,million
Platform	tax,jobs,health,federal,abortion
Scandal	accused,court,campaign,federal,commission

predicting each membership dummy using the features from the DTM. Table 1 presents the five words most predictive of each category—i.e., the five words with the largest coefficients in the ridge regression. As the table shows, the predictive words are highly sensible for all six categories. The campaign info category features generic words about candidates; the bio category features words about education, because candidate biographies almost always discuss candidates' educational backgrounds; the endorsement category by construction is based off of the word stem "endors"; the money category features words about money; the platform category features key policy words, like tax, health, and abortion; and the scandal category features words like accused and court, as would be expected.

Table 2 – Summary Statistics. Number of articles mentioning a candidate, by topic. Unit of observation is a candidate-year.

	Mean	SD	Min	Max	N
Total Articles Mentioning Candidate	3.33	7.26	0.00	156.00	4,747
Campaign Articles	2.08	4.87	0.00	118.00	4,747
Money Articles	0.31	1.29	0.00	36.00	4,747
Platform Articles	0.31	1.16	0.00	40.00	4,747
Scandal Articles	0.10	0.82	0.00	39.00	4,747
Biographical Articles	0.07	0.50	0.00	12.00	4,747
Endorsement Articles	0.45	1.82	0.00	60.00	4,747

3 Evaluating Primary Election Newspaper Coverage

We now turn to studying what features of primary elections local newspapers cover, using the dataset of newspaper coverage generated using the machine-learning procedure we just described.

3.1 How Much Newspaper Coverage of Primaries Is There?

We begin by offering some simple facts about the quantity of newspaper coverage in U.S. House primaries, focusing on the set of races for which we have access to newspaper data. Table 2 offers a summary of the data, where the unit of observation is a candidate in a given year's primary election in a given congressional district, for the set of contested U.S. House primaries. As the first row shows, on average, a candidate is mentioned in 3.33 newspaper articles over the course of the primary election—a modest level of overall coverage. This is consistent with what we know about House primaries from existing accounts; they are low information affairs with little coverage or polling.

The next rows break the coverage down by topic. As the second row shows, the majority of articles are classified as campaign coverage. As a reminder, this category includes articles that cover the basic nuts and bolts of campaigns—who is running, who has dropped out, and so forth. The remaining categories are all much rarer.

Figure 2 – Types of Primary Election News Coverage Over Time, Contested U.S. House Primaries, 1998–2012. Campaign coverage dominates newspaper coverage of primary elections.

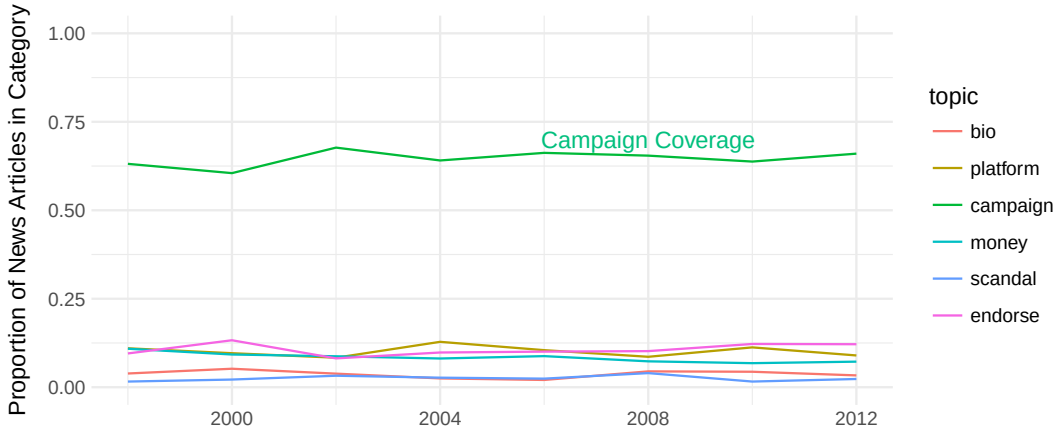


Figure 2 shows the proportion of each category in primary election news coverage over time. Campaign coverage consistently dominates newspaper coverage of House primaries. The other five categories are all much less common. Articles about candidate platforms are rare, never accounting for more than 8% of all articles in any given year. Articles containing news about endorsements are also relatively rare, accounting for roughly 1 out of every 10 articles in the sample.

3.2 Platform and Endorsement Coverage is Relatively Rare

Although aggregate rates of platform and endorsement articles are low, it is possible that only a minimal number of such articles are necessary in order to inform attentive readers. Accordingly, we also count the number of races in which we find zero articles about platforms and endorsements, respectively. Table 3 breaks shows the number of candidate years that receive some or no platform coverage and some or no endorsement coverage. As the upper left cell shows, in the majority of cases, 3,446 in all, candidates receive no newspaper coverage in the platform or endorsement categories. 324 candidates have at least one article mentioning an endorsement but have no articles about their platforms, while 312 have at least one article discussing their platforms but no articles about their endorsements. In only 77 cases do we

Table 3 – Rates of Platform and Endorsement Coverage. Presents the number of candidates who have, or do not have, any coverage of platforms and/or endorsements.

	No Platform Articles	One or More Platform Articles
No Endorsement Articles	3446	312
One or More Endorsement Articles	324	77

Unit of observation is a candidate-election.

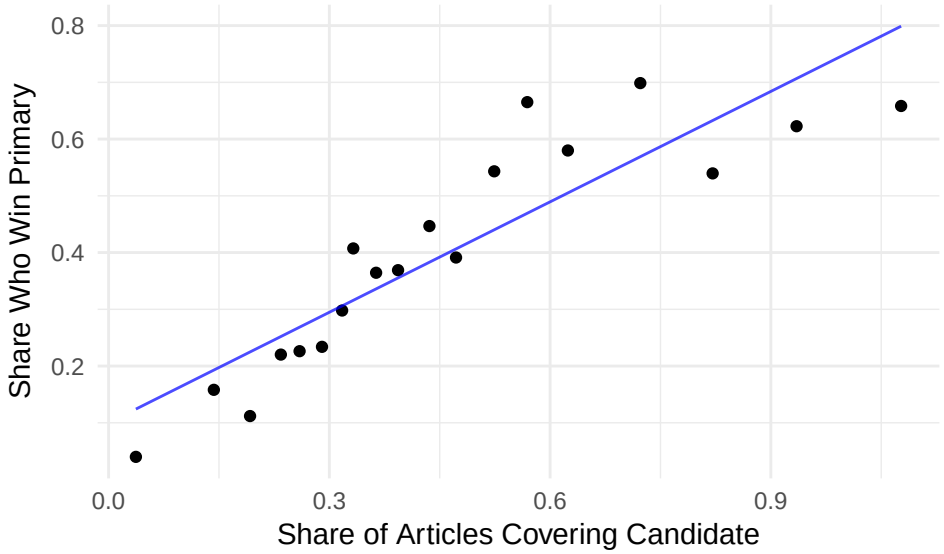
find both types of articles. Overall, only 17% of all candidate-years have any newspaper articles categorized as either platform or endorsement articles.

3.3 Newspaper Coverage Predicts Candidate Success

Although on average newspaper coverage appears to offer voters little direct information about candidate positions, it would be a mistake to conclude that it is altogether uninformative. In fact, receiving newspaper coverage is a strong predictor of candidate success in primary elections, which means that newsworthiness can be an informative signal for voters (note that there does not need to be a causal link between newspaper coverage and electoral performance for newspaper coverage to be an informative signal). Figure 3 plots average nomination rates against the average share of all articles in a primary that a particular candidate receives, in equal-sample-sized bins. As the plot shows, candidates who receive a larger share of newspaper coverage are substantially more likely to win nomination (R^2=0.29).

This relationship may be important because it provides information to voters in the absence of polls. In primaries with three or more candidates, there is a substantial risk of vote wasting (Duverger 1954). Strategically speaking, a voter should only vote for one of the top two candidates in the race, since a vote for any lesser candidate will have no impact on the winner. Hall and Snyder (2015*b*) shows that there is a meaningful amount of vote wasting in House primaries, and that this vote wasting goes down where media coverage is higher. The paper speculates that media information helps voters avoid vote wasting by conveying

Figure 3 – Primary Candidate News Coverage and Electoral Performance; Competitive Primaries in Open-Seat U.S. House Races. The share of news coverage devoted to a candidate predicts the candidate's success in the primary.



Note: Points are averages in equal-sample-sized bins of Log $ Raised. Regression line fit to underlying data. Generated using `binscatter` in Stata.

information about who the favored candidates are. Figure 3 supports this hypothesis; even in the absence of much polling of House elections, newsworthiness itself may help voters to identify who the favored candidates are. If a voter observes a candidate garnering a higher degree of newspaper coverage, she can infer that the candidate is expected to do well in the primary election.

It does not appear that newspaper coverage directly talks about electability; rather, simply the volume of news coverage is informative. To investigate this, we used a set of following words to look for articles that were informative of candidates' electability: "poll", "survey", "result", "percent", "win", "lead"/"led", "financial lead", "raise", "frontrunner", "underdog", "first place", "second place", and "runoff". We then closely read every article that contained any of these words to determine whether a given article actually contained information on the candidate's electability. Most electability articles were about polling results that showed which of the candidates were leading. For those who end up in a run-off,

articles on how they performed in primaries were also categorized as electability. We also selected articles containing nouns that were informative of the candidate's ranking, such as "frontrunner" or "first place." Of 3,282 Campaign Coverage articles, only 283 of them (8.6 percent) were informative about candidates' electability.

4 The Preference for Extremists and Newspaper Coverage

Thus far, we have shown that local newspapers offer relatively little coverage of House primaries, that little of this coverage contains information about candidate platforms, but that the level of coverage still provides useful information about candidates' electoral prospects. In this section, we turn to considering the possible effects this coverage has on the decisions that primary electorates make. The evidence we show is consistent with the idea that newspaper coverage informs primary voters about candidates and helps to undo the advantage of extremist candidates.

4.1 Scaling Candidates Using Campaign Contributions

Following Hall and Snyder (2015*a*), for each candidate i in primary election k we calculate her "relative extremism" compared to the other candidates in her primary as

$$\textit{Relative Extremism}_{ik} = |\textit{Cand Ideology}_{ik} - \textit{Most Moderate Ideology}_{k}|, \tag{1}$$

where $\textit{Most Moderate Ideology}_{k}$ is the estimated ideology of the most moderate candidate in primary k—that is, the rightmost candidate in a Democratic primary or the leftmost candidate in a Republican primary. By using the absolute distance from the most moderate candidate, we can pool over Democratic and Republican primaries. Candidates with higher values of Relative Extremism are farther into the wings of their respective parties than are

their primary opponents. As this definition hopefully makes clear, the idea of "extremism" here is entirely relative, and does not refer to any specific issue position or require any normative judgment on our part.

4.2 Extremists Outperform Moderates in U.S. House Primaries

We begin by comparing the performance of candidates with varying ideology in U.S. House races. To do so, we replicate and update the regressions estimated in Hall and Snyder (2015*a*). Specifically, we estimate equations of the form

$$Y_{ijkt} = \beta_0 + \beta_1 \, Relative\ Extremism_{ijkt} + \sum_i \beta_{2i} \, I(\#Cands_{jkt} = i) + \sum_{i=1}^{3} \beta_{3i} \left(Share\ of\ Donations_{ijkt}\right)^i + \sum_{i=1}^{3} \beta_{4i} \left(Share\ of\ Donors_{ijkt}\right)^i + \epsilon_{ijkt}, \quad (2)$$

where Y_{ijkt} is the vote share or an indicator for electoral victory for candidate i in party j's primary in district k at time t.

Following Hall and Snyder (2015*a*), we estimate this equation only using data for primary elections without an incumbent candidate, since incumbent candidates may be systematically more moderate and also more likely to win election. We also only estimate the equation for cases where at least two primary candidates receive scalings, since this is necessary to compute the measure of relative extremism. Finally, we re-scale the relative extremism measure so that it has mean 0 and standard deviation 1.

We also include controls for the share of donations and donors that candidate i receives in race party j's primary in district k at time t. Following Hall and Snyder (2015*a*), we include each of these controls as a flexible polynomial. The idea is to make comparisons among candidates with different estimated ideological positions but who raise similar amounts of money, to address the concern that candidates might look more extreme because they raise less money.

Table 4 – Extremists Outperform Moderates in U.S. House Primaries. These results are a replication, with more data, of those found in Hall and Snyder (2015*a*).

	Vote Pct	Vote Pct	Win	Win
Rel Extremism	0.17	0.35	0.03	0.03
	(0.13)	(0.13)	(0.00)	(0.00)
Controls	Yes	Yes	Yes	Yes
Controls Polynomial	1	3	1	3
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	5,836	5,836	5,836	5,836
# Elections	2,974	2,974	2,974	2,974

Vote Pct runs 0-100; Win is an indicator variable for winning the primary. Robust standard errors clustered by election in parentheses. Rel Extremism, defined in text, is standardized to have mean 0 and sd 1. Control variables are candidate's share of total donors in primary and candidate's share of total donations.

Table 4 presents the results, for both vote percentage and win probability, using two different specifications: one where the polynomials for the share of donations and of donors in equation 2 are first order, and one where they are third order (as in the equation above). Across all specifications, we see a positive association between relative extremism and electoral performance. A one standard deviation increase in relative extremism predicts a 0.21 or 0.38 percentage point increase in vote share, and a 3 percentage point increase in win probability. These associations are relatively modest, in size, but they are mostly precisely estimated.

The takeaway from this analysis is that extremist candidates tend to win U.S. House primary elections at somewhat higher rates than more moderate candidates. This relationship is purely descriptive and does not reflect the causal effect of a candidate *choosing* a more moderate or more extreme position. Indeed, there are many reasons extremists might do worse or better than moderates—we are only measuring the overall filtering of the primary candidate pool. Whatever the mechanism, this descriptive relationship tells us what types of ideology are represented in our legislatures. We now investigate whether this filtering

looks different in primary elections with more or less media coverage, to see whether more moderate or more extreme candidates do better when newspaper coverage is higher.

4.3 Extremist Advantage Concentrated in Low Newspaper-Coverage Elections

To see whether newspaper coverage changes the relationship between candidate ideology and success in primary elections, we use the Snyder and Stromberg (2010) measure of newspaper congruence. Districts with higher congruence are those where more of the newspaper's circulation is within the district, which leads the newspaper to devote more coverage to the district's member of Congress and its congressional elections. Snyder and Stromberg (2010) shows comprehensive evidence that higher congruence districts receive systematically more coverage of their members of Congress, and that voters in these districts are more informed about their members of Congress.[5] We scale this congruence measure to run from 0, in the least congruent district, to 1, in the most congruent district, and we re-estimate equations like those in Table 4 with an interaction between candidate ideology and congruence included.

The original congruence measure from Snyder and Stromberg (2010) only ran through 2004.[6] To extend it, we computed a new congruence measure using the formula outlined in Snyder and Stromberg (2010) with updated circulation-by-county figures from Alliance for Audited Media (AAM). We obtained county-circulation data for 2010 and 2011 for 286 (out of 2,039) local newspapers across 422 (out of 436) congressional districts in our dataset and back-filled the updated congruence measure for years 2003-2012.

The hope of using congruence is that it is an exogenous measure of newspaper coverage. Because congruence depends on the historical dispersion of newspapers and of readers, it may have little or nothing to do with the electoral features of present-day districts. However, we

[5]Many papers have used the congruence measure subsequently. As of this writing, Snyder and Stromberg (2010) has been cited nearly 500 times, according to Google scholar.

[6]Snyder and Stromberg (2010) analyze data from 1982 to 2004. They interpolated congruence data for the years 1983-1990, for which they did not have county-circulation data.

Table 5 – Extremist Primary Advantage Concentrated in Low News Coverage Elections.

	Vote Pct	Vote Pct	Vote Pct	Vote Pct
Rel Extremism	0.66	0.85	0.62	0.81
	(0.27)	(0.27)	(0.27)	(0.27)
Rel Extremism ×	-1.16	-1.21	-1.02	-1.07
Congruence	(0.51)	(0.51)	(0.52)	(0.52)
Congruence	1.02	1.02	1.30	1.23
	(0.62)	(0.63)	(0.81)	(0.81)
Candidate Controls	Yes	Yes	Yes	Yes
Cand Controls Polynomial	1	3	1	3
District Controls	No	No	Yes	Yes
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	4,585	4,585	4,585	4,585
# Elections	2,333	2,333	2,333	2,333

Vote Pct runs 0-100. Robust standard errors clustered by election in parentheses. Rel Extremism, defined in the text, is standardized to have mean 0 and sd 1; Congruence, also defined in the text, runs from 0 to 1, min to max. Candidate control variables are candidate's share of total donors in primary and candidate's share of total donations. District control variable are listed in text.

know that more congruent districts tend to be more rural, because urban areas have many districts served by a small number of large newspapers, and we might suspect there are other differences correlated with urban and rural areas. To account for this possibility, we follow Snyder and Stromberg (2010) and also estimate these regressions including a full set of variables about the districts as controls. Specifically, the controls are: percent urban in district; indicators for percent urban quintile; population density; indicators for density quintile; the number of congressional districts per city; log median income; percent senior citizens; percent military; percent farmer; percent foreign; and percent blue collar.

Table 5 presents the results. The pattern seems clear; the advantage of extremist candidates appears to be higher in low congruence areas, where newspaper coverage is more scant, and lower in more congruent places.[7] It is also somewhat encouraging that the coefficient

[7]We have also estimated these results using dummies for quartiles of the congruence variable, to ensure that our results are not driven by the strong assumption of linearity of the interaction of the two continuous variables (e.g., Hainmueller, Mummolo, and Xu 2018). Results are similar in this alternative setup.

estimates on this interaction variable do not change very much based on which controls we include. Moreover, the difference is large enough in magnitude that, in high congruence areas, the relationship inverts and we observe an advantage for more moderate candidates. In a hypothetical race with the highest level of congruence, we estimate that a one standard-deviation increase in relative extremism is associated with, in the smallest estimate (column 3), a 0.4 percentage-point decrease in vote share (-0.62-1.02 = -0.4). Although this relationship is not large in magnitude, it is in the opposite direction as conventional wisdom; more informed congressional primary electorates do not appear to favor more extreme primary candidates.

Perhaps because the advantage of extremist primary candidates in general is not large, we do not find a negative interaction of extremism and congruence when we use a binary indicator for victory as the outcome variable. However, the standard errors are very large, so that the confidence interval contains large negative or positive effects. The coarsening of the outcome variable evidently loses too much information for us to offer meaningful estimates on win probability given our sample size and statistical power.

4.4 More Congruent Areas Receive More Platform Coverage

Extremist candidates appear to perform worse in contested House primaries that occur in areas with more local newspaper coverage. Why might this be the case? In this subsection, we explore how the nature of coverage differs in places with more congruent news coverage, and we find that, in high congruence areas, local newspaper articles offer more information about candidate platforms.

Specifically, we estimate equations of the form

$$\#\ Platform\ Articles_{it} = \beta_1 Congruence_{it} + \sum_i \beta_{2i} I(\#Cands_{it} = i) + X_{it} + \epsilon_{it}, \qquad (3)$$

Table 6 – More Platform Information Where Newspaper Congruence is Higher.

	# of Articles About Platforms	
Congruence	0.32	0.33
	(0.09)	(0.10)
Controls	No	Yes
# Cand Fixed Effects	Yes	Yes
# Observations	2,410	2,410
# Elections	718	718

Congruence runs from 0 to 1, min to max. Robust standard errors clustered by election in parentheses.

where X_{it} is an optional vector of control variables. Like before, we use this vector to attempt to control for potential differences between high congruence and low congruence districts.

Table 6 presents the results. As the table shows, higher congruence areas appear to receive more newspaper articles about candidate platforms. The second column presents the estimates with the inclusion of the full suite of district control variables, finding very similar results. The results suggest that extremist candidates may do worse in more congruent areas in part because voters in these areas have more information about candidate platforms, though there are many steps along that causal chain that we cannot observe in our data.

4.5 Newsworthiness More Informative in More Congruent Areas

Another possibility, not mutually exclusive with the increase in platform coverage, is that newsworthiness is also more informative in high-information areas, because there are more articles in general. If more moderate candidates receive a higher proportion of news coverage, then the increased informedness of news coverage in high congruence areas could help to explain the reduced advantage of more extreme candidates in higher congruence areas. To test this, we run regressions predicting vote share and relative extremism, respectively, as

Table 7 – News Coverage is More Informative in Higher Congruence Areas.

	Vote Pct	Vote Pct	Rel Extremism	Rel Extremism
Article Share	22.63	22.43	0.04	0.09
	(2.79)	(2.80)	(0.22)	(0.22)
Article Share ×	14.66	15.07	-0.57	-0.59
Congruence	(5.59)	(5.70)	(0.43)	(0.43)
Congruence	-6.96	-6.07	0.26	0.28
	(2.08)	(2.22)	(0.29)	(0.30)
District Controls	No	Yes	No	Yes
# Cand Fixed Effects	Yes	Yes	Yes	Yes
# Observations	1,512	1,512	1,076	1,076
# Elections	718	718	606	606

Vote Pct runs 0-100, as does Article Share. Robust standard errors clustered by election in parentheses. Rel Extremism and Congruence, both defined in text, are standardized to have mean 0 and sd 1. District control variable are listed in text.

a function of a candidate's article share, interacting article share with congruence as well. Table 7 presents the results.

As the first two columns show, the relationship between a candidate's share of articles in a primary—that is, her relative newsworthiness—is strongly associated primary vote share, even in low congruence places (first row). As the second row shows, this relationship is considerably larger in high congruence areas. This is true with or without district controls. This suggests that newsworthiness is an especially good leading indicator of primary success in high-congruence areas.

This relationship may help to explain the diminished advantage of more extreme candidates in high congruence areas if newsworthiness helps primary voters to pick out more moderate candidates. The second two columns suggest this may be the case, but the results are too imprecise to draw any strong conclusions. In these columns, we see that, in low congruence areas, there is no apparent relationship between a candidate's article share and her relative ideological extremism. However, in higher congruence areas, there is a negative

though statistically imprecise relationship—that is, candidates in high congruence areas who receive a higher article share appear to be more moderate, on average, consistent with the possibility that newsworthiness is a signal both of electability and lower extremism.

5 Conclusion

The role of primary electorates in the polarization of American politics is much disputed, with scholars debating whether primary electorates are more extreme than other voters, and debating if primary elections encourage polarization or not. In this paper, we have taken a different approach to studying this question. Rather than attempting to measure the issue preferences of primary voters, we have examined how their behavior in real elections varies along with the information environment. Using a new dataset of local newspaper articles in U.S. House primary elections, we have shown that there is relatively little news coverage of primary elections—and of what coverage there is, very little of it concerns candidate platforms.

However, there is important variation in how much information primary voters receive. Where newspaper coverage is higher, more extreme candidates do worse in contested House primaries. Moreover, in these areas, newspaper coverage offers more information about candidate platforms. Together, the evidence suggests that information about candidate platforms may influence the choices that primary voters make, in the direction opposite what conventional wisdom would predict. Instead of helping primary voters to pick out extreme candidates, newspaper information about candidate platforms may encourage them to pick more moderate nominees.

The decline of local news coverage is an important story in American politics. Although our evidence does not directly estimate the causal effect of the decline in the capacity of local news coverage, it certainly suggests that further declines may lead to an increasing

advantage for more extreme primary candidates. At the very least, our results suggest that this is a phenomenon that warrants further study in the future.

References

Ansolabehere, Stephen, John Mark Hansen, Shigeo Hirano, and James M. Snyder Jr. 2010. "More Democracy: The Direct Primary and Competition in US Elections." *Studies in American Political Development* 24(2): 190–205.

Boatright, Robert G. 2013. *Getting Primaried: The Changing Politics of Congressional Primary Challenges*. University of Michigan Press.

Brady, David W., Hahrie Han, and Jeremy C. Pope. 2007. "Primary Elections and Candidate Ideology: Out of Step with the Primary Electorate?" *Legislative Studies Quarterly* 32(1): 79–105.

Duverger, Maurice. 1954. *Political Parties: Their Organization and Activity in the Modern State*. New York: Wiley.

Fowler, Anthony, and Michele Margolis. 2014. "The Political Consequences of Uninformed Voters." *Electoral Studies* 34: 100–110.

Gilliam, Jr., Franklin D., and Shanto Iyengar. 2000. "Prime Suspects: The Influence of Local Television News on the Viewing Public." *American Journal of Political Science* 44(3): 560–573.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267?297.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing. Xu. 2018. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* .

Hall, Andrew B., and Daniel M. Thompson. 2018. "Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in U .S. Elections." *American Political Science Review* .

Hall, Andrew B., and James M. Snyder, Jr. 2015*a*. "Candidate Ideology and Electoral Success." Working Paper.

Hall, Andrew B., and James M. Snyder, Jr. 2015*b*. "Information and Wasted Votes: A Study of U.S. Primary Elections." *Quarterly Journal of Political Science* 10(4): 433–459.

Hill, Seth J., and Chris Tausanovitch. 2017. "Southern Realignment, Party Sorting, and the Polarization of American Primary Electorates, 1958-2012." *Public Choice* .

Hirano, Shigeo, James M. Snyder, Jr., Stephen Ansolabehere, and John Mark Hansen. 2010. "Primary Elections and Partisan Polarization in the US Congress." *Quarterly Journal of Political Science* 5(2): 169–191.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2): 254–277.

Martin, Gregory J., and Josh McCrain. 2018. "Local News and National Politics." Working Paper.

McCarty, Nolan M., Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press Cambridge, MA.

McGhee, Eric, Seth Masket, Boris Shor, Steven Rogers, and Nolan McCarty. 2014. "A Primary Cause of Partisanship? Nomination Systems and Legislator Ideology." *American Journal of Political Science* 58(2): 337–351.

Owen, Guillermo, and Bernard Grofman. 2006. "Two-stage Electoral Competition in Two-Party Contests: Persistent Divergence of Party Positions." *Social Choice and Welfare* 26(3): 547–569.

Peterson, Erik. 2017. "The Role of the Information Environment in Partisan Voting." *The Journal of Politics* 79(4): 1191–1204.

Peterson, Erik. 2018. "Paper Cuts: How Reporting Resources Affect Political News Coverage." Working Paper.

Pildes, Richard H. 2011. "Why the Center Does Not Hold: the Causes of Hyperpolarized Democracy in America." *California Law Review* pp. 273–333.

Riggle, Ellen D. 1992. "Cognitive Strategies and Candidate Evaluations." *American Politics Quarterly* 20(2): 227–246.

Sides, John, Chris Tausanovitch, Lynn Vavreck, and Christopher Warshaw. 2017. "On the Representativeness of Primary Electorates." *British Journal of Political Science* .

Snyder, Jr., James M., and David Stromberg. 2010. "Press Coverage and Political Accountability." *Journal of Political Economy* 118(2): 355–408.

Thomsen, Danielle M. 2018. "When Might Moderates Win the Primary?" In *Routledge Handbook of Primary Elections*, ed. Robert G. Boatright. Routledge pp. 226–235.

Online Appendix

Intended for online publication only.

Appendix A. Categorization Scheme

Election news coverage during contested House primaries are categorized into one of six categories. Here, we describe in detail how the categories are defined, along with examples.

1. Campaign Coverage: information on who has entered or dropped out of the race; overview of who is running against whom; a report on election or polling results; prediction on who's going to win based on polling results; information on campaign activities (e.g., ads, commercials, meeting with voters); backing from a former opponent who is no longer running. However, if a candidate receives endorsement or is backed by an interest group with a clear policy or ideological stance, the article is categorized as Platform. Likewise, if an article describes in detail the issues that were discussed in advertisements or during campaign events, it goes under Platform.

• State Senate majority leader Garagiola 6th District candidate: CUMBERLAND Rob Garagiola, Democratic candidate for Marylands 6th Congressional District, knocked on some doors in the Mapleside area Saturday in his quest to win the upcoming primary election and then unseat U.S. Rep. Roscoe Bartlett.Im getting a lot of positive feedback. Some people are not aware of the primary election which is a month away today. Were taking one step at a time, said Garagiola, the majority leader in Marylands state Senate and one of four Democratic candidates...

• Reed drops out of 6th District race: Dr. Maureen Reed has dropped out of the 6th District U.S. House contest, saying she's stepping aside in order to focus the race on beating incumbent Rep. Michele Bachmann."During the past few days, I have come to the conclusion that a prolonged primary fight only assists Michele Bachmann," Reed wrote on her campaign blog. "I feel that it is time for the DFL to unify behind one candidate in this race."Reed, 57 a resident of the Grant, had been the Independence Party candidate for...

• Poll good news for Brown: Campaign Bits By Tom Waring Polls released on Friday by the Center for Opinion Research show Melissa Brown leading a three-way Republican battle in the 13th Congressional District, while the two Democrats are in a close race.Among Republicans likely to vote in the April 27 primary, Brown has 36 percent of the vote. State Rep. Ellen Bard follows with 20 percent. Al Taubenberger trails with 11 percent."This poll shows we are in excellent position for a victory in April...

2. Candidate Biography: Biographical information (name, age, past experience, etc.) about a candidate; interviews; section in a newspaper that is specifically dedicated for candidate running in a congressional district in which the newspaper is published.

• Christopher Brent Reilly: Age: 50.He lives: York Township, York County. Education: B.A. in government and politics, University of Maryland.Family: Wife, Lisa, and three children, Patrick, William and Claire.Occupation: York County Commissioner.Hobbies: Fishing; reading; and cooking ethnic food.First job: Shoeshine boy at a barbershop.Attribute/ability

he will take to Washington: Experience. I have a proven record of fiscal conservatism. I'm a Conservative and I'm...

• Candidate Q&A - Jonathan Paton, candidate for Congressional District 1: Name: Jonathan Paton Age: 41 Occupation: Self, Public Relations Where do you live? Pima County How long have you lived in the area? My entire life.Short description about yourself (200 words) After serving for two years as a Representative in the state Legislature, I felt called to serve my country in the darkest days of the war. So I voluntarily enlisted to serve in Iraq on the front lines as an operations officer in an intelligence unit. The experience...

• Congressional candidate Dan Roberts rejected family's politics; wounded in Vietnam: Editor's note: This is one in a series about the 2nd District congressional candidates.By Richard Halstead Marine Corps 2nd Lt. Dan Roberts was leading his troops back from a patrol through Vietnam's Elephant Valley near Da Nang in 1966 when a mine exploded, killing one soldier and piercing Roberts' left leg with shrapnel. "I guess there was an element of shock and disbelief. I had these fragments of shrapnel going through my left calf," Roberts said. After his radio...

3. Money: Candidate's fundraising efforts or financial disclosure. "Money" articles must involve indicators of financial support or money. For instance, if an article is about an interest group endorsing a candidate but it does not mention financial support, the article is categorized under Platform. Similarly, if an article is primarily about a candidate discussing policy issues during a fundraiser and does not mention how much money the candidate raised, the article is categorized under Platform.

• DUNCAN WAR CHEST NOW AT $200,000: State Sen. Jim Duncan has made a strong financial start in his campaign to unseat Republican incumbent U.S. Rep. Don Young in November, reports filed with the Federal Election Commission show. Duncan has raised more than $200,000, with four months to go until his first election test, the August Democratic primary. Money counts in a statewide race, and at his current fund-raising pace, Duncan could become the strongest challenger Young has faced in years.Young has collected three...

• U.S. House candidates Daines, Smith, Rankin report more than $1M in assets: Diane Benson, Democratic candidate for Congress and mother of an Iraq war veteran who lost his legs there, said Friday that the decision to extend the tour of the Alaska-based 172nd Stryker Brigade is wrong.Benson, in a prepared statement, said U.S. Rep. Don Young should be speaking out about it."I call on Rep. Don Young to immediately stand up for our Alaskan sons and daughters and demand that our Alaskan families are reunited as planned. We cannot allow our families to suffer...

• Texas-based Super PAC Campaign for Primary Accountability targets US Rep. Spencer Bachus, backs challenger Scott Beason: WASHINGTON – A Texas-based political action committee with $1.6 million cash on hand will be spending some of that money to help defeat U.S. Rep. Spencer Bachus, a 10-term veteran who PAC organizers have targeted because of his longevity and ethics investigation."Incumbents like Mr. Bachus ... are longtime passengers on the inside-the-beltway gravy train," said Curtis Ellis, a spokesman for the Campaign for Primary Accountability.The entrance of a Super PAC – which can spend...

• Bellavia turns financial disclosure into prodding of Collins: which, in Bellavia's case, is minimal.Bellavia this week released the personal financial disclosure statement he is required

to file for his candidacy for the Republican nomination for Congress in New York's 27th district.And the document shows the Iraq War hero with family income so far this year of no more than $11,820, along with a credit card debt somewhere between $15,001 and $50,000.

4. Platform: Information on candidates' policy stance or ideological platform.

• BENTON'S ABORTION VIEWS CHANGE: As he wades ashore in the battle to capture his partys congressional nomination, state Sen. Don Benton has made a major shift in his position on abortion rights.Like the three other Republican soldiers battling for this beach, Benton is now in the pro-life or anti-choice camp.He had been the only pro-choice candidate among the four GOP contenders for the office. The others, Pat Fiske, Paul Phillips and Rick Jackson, were already dug in as opponents of general legalized abortion in...

• Harris wants state gas tax to be suspended; Kratovil calls suggestion irresponsible: The Republican candidate for Maryland's 1st congressional district wants the state gas tax suspended for three months.His Democratic opponent said suspending the gas tax would be irresponsible without coming up with a plan to offset the loss of tax revenues to the state.State Sen. Dr. Andy Harris, R-7th, Baltimore and Harford counties, said Gov. Martin O'Malley should call a one-day special session of the legislature so lawmakers can suspend the state tax on gasoline and diesel...

5. Scandal: Any case in which a candidate's ethics is called into question; investigation; lawsuit; legal dispute; allegation.

• Griffith files suit against campaign manager: Defeated candidate claims some funds not accounted for Parker Griffith, who was defeated in the March 13 Republican primary in his bid to return to Congress, filed a lawsuit Thursday against his former campaign manager, alleging breach of contract and failure to properly account for campaign funds.The suit filed in Madison County Circuit Court contends Griffith hired Huntsville resident Barbara Nash on Jan. 12 to work as his campaign manager in the 5th Congressional District race...

• Paton accuses primary foe of fraud, seeks removal from congressional ballot in Arizona: As the Democratic primary to nominate the replacement for Congressman Mike Ross unfolds, it appears that none of the three Democratic candidates comes close to filling his large shoes.The fundraising reports reveal that Hot Springs attorney Q. Byrum Hurst has the most backing of those willing to donate money. But the more we learn about his background, the more you have to wonder why. Hurst reported raising over $100,000 in the first month of his campaign. His opponents State...

• Columbia cops arrest state representative for DUI, weapons possession, Ted Vick is a candidate in race for a new congressional seat: Columbia police officers arrested a state representative Thursday for driving under the influence of alcohol and the unlawful carrying of a pistol after he was stopped for speeding. S.C. Rep. Ted Vick, D-Chesterfield, was released from the Alvin S. Glenn Detention Center on personal recognizance bonds for the charges. He also was given a ticket for speeding.Vick, 39, is one of several candidates seeking the Democratic nomination for South Carolinas new 7th congressional seat....

6. Endorsement: Any article in which the word stem "endors" is found.

Appendix B: Performance of Various Supervised Learning Algorithms

Table A.1 – Precision, Recall, and F-scores for each classification algorithm.

Algorithm	Precision	Recall	F-score
SVM	0.798	0.696	0.734
SLDA	0.780	0.682	0.722
Maximum Entropy	0.702	0.702	0.698
Boosting	0.714	0.626	0.658
Random Forest	0.894	0.592	0.656
Bagging	0.784	0.558	0.602
Neural Network	0.552	0.514	0.520
Tree	0.474	0.486	0.472